03-27-06

AF/1646

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In application of:	Examiner: Kemmerer, Elizabeth
Avi J. ASHKENAZI, et al.	Art Unit: 1646
Application Serial No. 09/978,295	Confirmation No. 6495
Filed: October 15, 2001	Attorney's Docket No. 39780-2630 P1C11
For: **SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS ENCODING THE SAME**	**Customer No. 35489**

EXPRESS MAIL LABEL NO. EV 582 620 470 US
DATE MAILED: MARCH 24, 2006

ON APPEAL TO THE BOARD OF PATENT APPEALS AND INTERFERENCES

APPELLANTS' BRIEF

MAIL STOP APPEAL BRIEF - PATENTS
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

Dear Sir:

On August 29, 2005, the Examiner made a final rejection to pending Claims 58-62. A Notice of Appeal was filed on January 25, 2006.

Appellants hereby appeal to the Board of Patent Appeals and Interferences from the last decision of the Examiner.

The following constitutes Appellants' Brief on Appeal.

PTO/SB/17 (01-03)
Approved for use through 04/30/2003. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

FEE TRANSMITTAL for FY 2005

Effective 12/08/2004. Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT	($)500.00

Complete if Known	
Application Number	09/978,295
Filing Date	October 15, 2001
First Named Inventor	Avi J. Ashkenazi
Examiner Name	Kemmerer, Elizabeth.
Art Unit	1646
Attorney Docket No.	39780-2630P1C11

METHOD OF PAYMENT (check one)

☐ Check ☐ Credit card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:

Deposit Account Number: 08-1641 (Docket No. 39780-2630P1C11)

Deposit Account Name: Heller Ehrman LLP

The Commissioner is authorized to: *(check all that apply)*

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) during the pendency of this application

☐ Charge fee(s) indicated below, **except for the filing fee** to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Fee Code	Entity Fee ($)	Small Fee Code	Entity Fee ($)	Fee Description	Fee Paid
1001	300	2001	150	Utility filing fee	
1002	350	2002	175	Design filing fee	
1003	550	2003	275	Plant filing fee	
1004	790	2004	395	Reissue filing fee	
1005	200	2005	100	Provisional filing fee	

SUBTOTAL (1) ($)

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Extra Claims		Fee from below		Fee Paid
Total Claims		-20** =		x		=	
Independent Claims		-3** =		x		=	0
Multiple Dependent						=	0

Large Fee Code	Entity Fee ($)	Small Fee Code	Entity Fee ($)	Fee Description
1202	50	2202	25	Claims in excess of 20
1201	200	2201	100	Independent claims in excess of 3
1203	360	2203	180	Multiple dependent claim, if not paid
1204	200	2204	100	**Reissue independent claims over original patent
1205	50	2205	25	**Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($)

***or number previously paid, if greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Fee Code	Entity Fee ($)	Small Fee Code	Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for *ex parte* reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	120	2251	60	Extension for reply within first month	
1252	450	2252	225	Extension for reply within second month	
1253	1,020	2253	510	Extension for reply within third month	
1254	1,590	2254	795	Extension for reply within fourth month	
1255	2,160	2255	1,080	Extension for reply within fifth month	
1401	500	2401	250	Notice of Appeal	
1402	500	2402	250	Filing a brief in support of an appeal	500.00
1403	1,000	2403	500	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	500	2452	250	Petition to revive - unavoidable	
1453	1,500	2453	750	Petition to revive - unintentional	
1501	1,400	2501	700	Utility issue fee (or reissue)	
1502	800	2502	400	Design issue fee	
1503	1,100	2503	550	Plant issue fee	
1460		1460		Petitions to the Commissioner	
1807	50	1807	50	Processing fee under 37 CFR 1.17(q)	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	790	2809	395	Filing a submission after final rejection (37 CFR 1.129(a))	
1810	790	2810	395	For each additional invention to be examined (37 CFR 1.129(b))	
1801	790	2801	395	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

* Reduced by Basic Filing Fee Paid SUBTOTAL (3) $500.00

SUBMITTED BY

Complete *(if applicable)*

Name *(Print/Type)*	Barrie D. Greene	Registration No. *(Attorney/Agent)*	46,740	*Telephone*	650-324-7000
Signature		*Date*	March 24, 2006	Customer No.	35489

Please type a plus sign (+) inside this box [+]

PTO/SB/21 (6-99)
Approved for use through 09/30/2000. OMB 0651-0031
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

TRANSMITTAL FORM

(to be used for all correspondence after initial filing)

Application Number	09/978,295
Filing Date	October 15, 2001
First Named Inventor	Avi J. Ashkenazi
Group/Art Unit	1646
Examiner Name	Kemmerer, Elizabeth
Total Number of Pages in This Submission	
Attorney Docket Number	39780-2630P1C11

ENCLOSURES ***(check all that apply)***

- [X] **Fee Transmittal Form**
  - [X] **Fee Attached**
- [ ] AMENDMENT / RESPONSE
  - [ ] After Final
  - [ ] Version With Markings Showing Changes
  - [ ] Affidavits/declaration(s)
- [ ] Extension of Time Request
- [ ] INFORMATION DISCLOSURE STATEMENT WITH FORM PTO-1449
- [ ] Certified Copy of Priority Document(s)
- [ ] Response to Missing Parts/ Incomplete Application
  - [ ] Response to Missing Parts under 37 CFR 1.52 or 1.53
  - [ ] Copy of Notice
- [ ] DECLARATION
- [ ] DECLARATION ON INCORPORATION BY REFERENCE
- [ ] AMENDMENT UNDER 37 CFR §1.48(b)
- [ ] Petition Routing Slip (PTO/SB/69) and Accompanying Petition
- [ ] Petition to Convert to a Provisional Application
- [ ] Power of Attorney, by Assignee to Exclusion of Inventor Under 37 C.F.R. §3.71 With Revocation of Prior Powers
- [ ] Terminal Disclaimer
- [ ] Small Entity Statement
- [ ] Request for Refund
- [ ] After Allowance Communication to Group
- [ ] Appeal Communication to Board of Appeals and Interferences
- [X] **Appeal Communication to Group** ***(Appeal Notice, Brief, Reply Brief)***
- [ ] Proprietary Information
- [ ] Status Letter
- [X] **ADDITIONAL ENCLOSURE(S)** ***(PLEASE IDENTIFY BELOW):***
- [X] **EVIDENCE APPENDIX ITEMS1-18; STAMPED RETURN POSTCARD**

Remarks

AUTHORIZATION TO CHARGE DEPOSIT ACCOUNT 08-1641 FOR ANY FEES DUE IN CONNECTION WITH THIS PAPER, REFERENCING ATTORNEY'S DOCKET NO. 39780-2630 P1C11.

SIGNATURE OF APPLICANT, ATTORNEY OR AGENT

Firm or Individual name	**HELLER EHRMAN LLP**	**Barrie D. Greene (Reg. No. 46, 740)**	
	275 Middlefield Road, Menlo Park, California 94025	**Telephone: (650) 324-7000**	**Facsimile: (650) 324-0638**
Signature			
Date	**March 24, 2006**	Customer Number:	**35489**

CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. §1.10 on the date indicated below and addressed to: **MAIL STOP APPEAL BRIEFS-PATENTS,** Commissioner for Patents, PO Box 1450, Alexandria, Virginia 22313-1450, on this date: **March 24, 2006.**

Express Mail Label **EV 582 620 470 US**

Typed or printed name	**ELENA TORRES**		
Signature		Date	**March 24, 2006**

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Mail Stop APPEAL BRIEFS-PATENT, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

1. REAL PARTY IN INTEREST

The real party in interest is Genentech, Inc., South San Francisco, California, by an assignment of the patent application U.S. Serial No. 09/918,585 recorded July 30, 2001, at Reel 012095 and Frame 0677.

2. RELATED APPEALS AND INTERFERENCES

There are no related appeals or interferences known to Appellants, Appellants' legal representative, or Appellants' assignee that will directly affect or be directly affected by or have a bearing on the Board's decision in the present appeal.

3. STATUS OF CLAIMS

Claims 58-62 are in this application.

Claims 1-57 and 63 are canceled.

Claims 58-62 stand rejected and Appellants appeal the rejection of these claims.

A copy of the rejected claims involved in the present Appeal is provided in the Claims Appendix.

4. STATUS OF AMENDMENTS

There were no amendments to the claims submitted after final rejection. All previous amendments to the claims have been entered.

5. SUMMARY OF CLAIMED SUBJECT MATTER

The invention claimed in the present application concerns an isolated antibody that specifically binds to the polypeptide of SEQ ID NO:132 (Claim 58). The invention further provides monoclonal antibodies (Claim 59), humanized antibodies (Claim 60), antibody fragments (Claim 61), and labeled antibodies (Claim 62) that specifically bind to the polypeptide of SEQ ID NO:132.

Support for the preparation and uses of antibodies is found throughout the specification, including, for example, pages 217-225. The preparation of antibodies is described in Example

104, while Example 106 describes the use of the antibodies for purifying the polypeptides to which they bind. Isolated antibodies are defined in the specification at page 132, lines 29-38. Support for monoclonal antibodies is found in the specification at, for example, page 217, line 30, to page 219, line 11, and Example 104. Support for humanized antibodies is found in the specification at, for example, page 219, line 12, to page 220, line 14. Support for antibody fragments is found in the specification at, for example, page 131, line 29, to page 132, line 22, and page 221, lines 6-34. Support for labeled antibodies is found in the specification at, for example, page 133, lines 1-4, and page 224, line 35, to page 225, line 4.

The polypeptide of SEQ ID NO:132 is designated PRO351, and its amino acid sequence is shown in Figure 49, while the encoding nucleic acid sequence (SEQ ID NO:131) is shown in Figure 48. Page 112 line 37 through page 113, line 2 of the specification provides the description for Figures 48 and 49. The specification discloses that various portions of the PRO351 polypeptide possess significant sequence similarity to the serine protease prostasin (see, for example, page 12, lines 21-33). The isolation of cDNA clones encoding PRO351 of SEQ ID NO:7 is described in Example 22. Examples 100-103 describe the expression of PRO polypeptides in various host cells, including *E. coli*, mammalian cells, yeast and Baculovirus-infected insect cells. Finally, Example 114, in the specification at page 331, line 23, to page 346, line 4, sets forth a Gene Amplification assay which shows that the PRO351 gene is amplified in the genome of certain human lung cancers (see Table 9).

The specification discloses that antibodies to PRO polypeptides may be used, for example, in purification of PRO (page 225, lines 5-11 and Example 106), in diagnostic assays for PRO expression (page 190, lines 3-9, and page 224, line 21 to page 225, line 4), as antagonists to PRO (page 198, lines 3-6), and as elements of pharmaceutical compositions for the treatment of various disorders (page 223, line 30, to page 224, line 28).

6. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

I. Whether Claims 58-62 satisfy the utility requirement of 35 U.S.C. §101.

II. Whether Claims 58-62 satisfy the enablement requirement of 35 U.S.C. §112, first paragraph.

7. ARGUMENT

Summary of the Arguments:

Issue I: Utility

Patentable utility of the PRO351 polypeptide and the antibodies which bind it is based upon the gene amplification data for the gene encoding the PRO351 polypeptide. The specification discloses that the gene encoding PRO351 showed significant amplification, ranging from 2.03 to 2.75- fold, in ten different lung primary tumors. The Declaration of Dr. Audrey Goddard, submitted with Appellants' Response filed April 29, 2004, explains that a gene identified as being amplified at least 2-fold by the disclosed gene amplification assay in a tumor sample relative to a normal sample is useful as a marker for the diagnosis of cancer, for monitoring cancer development and/or for measuring the efficacy of cancer therapy. Accordingly, the Examiner's assertion that "the specification provides data showing a very small increase in DNA copy number, approximately 2-fold, in a few tumor samples for PRO351" (Page 2 of the Advisory Action mailed October 14, 2004) is both factually and scientifically incorrect. By referring to the 2.0-fold to 3.1-fold amplification of the PRO274 gene in lung tumors as "very small," the Examiner ignores the teachings of an expert declaration without any basis, or without presenting any evidence to the contrary.

The Examiner has asserted that the control sample used in the disclosed gene amplification experiments was not a proper control, stating that "the art does not consider pooled, unrelated DNA samples to be an appropriate control." (Page 5 of the Office Action mailed August 29, 2005). Appellants submit that the negative control taught in the specification was known in the art at the time of filing, and accepted as a true negative control as demonstrated by use in peer reviewed publications.

The Examiner has asserted that "damaged, precancerous lung epithelium is often aneuploid," and stated that "[o]ne skilled in the art would not conclude that PRO351 is a diagnostic probe for lung cancer unless it is clear that PRO351 is amplified to a clearly greater extent in true lung tumor tissue relative to non-cancerous lung epithelium." (Page 4 of the Office Action mailed August 29, 2005). In support of this assertion the Examiner cited a reference by Hittelman. Hittelman actually shows that an increase in chromosome number or gene

amplification is associated not with normal tissues, but with cancerous, or pre-cancerous tissues, and therefore, an increase in chromosome number or gene amplification is a useful marker for a cancerous or pre-cancerous state.

The Examiner has asserted that the disclosed gene amplification data does not establish a patentable utility for the PRO351 polypeptide or the claimed antibodies that bind it because allegedly "[t]he art demonstrates that gene amplification does not reliably correlate with increased mRNA transcript levels or increased polypeptide levels," citing references by Pennica *et al.* and Konopka *et al.* (Page 4 of the Office Action mailed February 4, 2004). Haynes *et al.* was cited "as providing evidence that polypeptide expression levels cannot be accurately predicted from mRNA levels" (Page 3 of the Advisory Action mailed October 14, 2004). In further support of the alleged lack of correlation between mRNA levels and polypeptide levels, the Examiner has cited additional references by Hu *et al.*, LaBaer, Chen *et al*, Lian *et al.*, Fessler *et al.*, Gygi *et al.* and Greenbaum *et al.*

The Examiner's reference to the lack of necessary correlation or accurate prediction in some of the rejections (as Appellants will discuss in the detailed arguments) clearly shows that the Examiner applied an improper legal standard when making this rejection. The evidentiary standard to be used throughout *ex parte* examination in setting forth a rejection is a preponderance of the totality of the evidence under consideration. Thus, to overcome the presumption of truth that an assertion of utility by the Appellant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Only after the Examiner has made a proper *prima facie* showing of lack of utility, does the burden of rebuttal shift to the Appellant. The references cited by the Examiner do not suffice to make a *prima facie* case that more likely than not no generalized correlation exists between gene (DNA) amplification and increased mRNA and polypeptide levels.

In contrast, Appellants have submitted ample evidence to show that, in general, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level. First, the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* (made of record in Appellants' Response filed April 29, 2004) collectively teach that in general, gene

amplification increases mRNA expression. Second, the Declaration of Dr. Paul Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, shows that, in general, there is a correlation between mRNA levels and polypeptide levels. Appellants further note that the sale of gene expression chips to measure mRNA levels is a highly successful business, with a company such as Affymetrix recording 168.3 million dollars in sales of their GeneChip arrays in 2004. Clearly, the research community believes that the information obtained from these chips is useful (i.e., that it is more likely than not informative of the protein level).

Taken together, although there are some examples in the scientific art that do not fit within the central dogma of molecular biology that there is a correlation between DNA, mRNA, and polypeptide levels, these instances are exceptions rather than the rule. In the majority of amplified genes, as exemplified by Orntoft *et al.*, Hyman *et al.*, Pollack *et al.*, and the Polakis Declaration, the teachings in the art overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO351 gene, that the PRO351 polypeptide is concomitantly overexpressed. Thus, the claimed antibodies that bind the PRO351 polypeptide have utility in the diagnosis of cancer.

Even if there is no correlation between gene amplification and increased mRNA/protein expression, (which Appellants expressly do not concede), a polypeptide encoded by a gene that is amplified in cancer would still have a specific, substantial, and credible utility. As evidenced by the Ashkenazi Declaration and the teachings of Hanna and Mornin, simultaneous testing of gene amplification and gene product over-expression enables more accurate tumor classification, even if the gene-product, the protein, is not over-expressed. This leads to better determination of a suitable therapy for the tumor, as demonstrated by the real-world example of the breast cancer marker HER-2/neu.

Accordingly, Appellants submit that when the proper legal standard is applied, one should reach the conclusion that the present application discloses at least one patentable utility for the for the PRO351 polypeptide and the claimed antibodies which bind it.

Issue II: Enablement

Claims 58-62 stand rejected under 35 U.S.C. §112, first paragraph, allegedly "since the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention." (Pages 2-3 of the Office Action mailed August 29, 2005).

Appellants submit that, as discussed above, the PRO351 polypeptide and the antibodies that bind it have utility in the diagnosis of cancer. Based on such a utility, one of skill in the art would know exactly how to use the claimed antibodies for diagnosis of cancer, without any undue experimentation.

These arguments are all discussed in further detail below under the appropriate headings.

ISSUE I: Claims 58-62 satisfy the utility requirement of 35 U.S.C. §101

Claims 58-62 stand rejected under 35 U.S.C. §101 because allegedly "the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility." (Page 2 of the Office Action mailed August 29, 2005).

Appellants submit, for the reasons set forth below, that the specification discloses at least one credible, substantial and specific asserted utility for the claimed antibodies that bind the PRO351 polypeptide.

A. The Legal Standard for Utility

According to 35 U.S.C. §101:

> Whoever invents or discovers any new and *useful* process, machine, manufacture, or composition of matter, or any new and *useful* improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title. (Emphasis added.)

In interpreting the utility requirement, in *Brenner v. Manson*[1] the Supreme Court held that the *quid pro quo* contemplated by the U.S. Constitution between the public interest and the interest of the inventors required that a patent Appellant disclose a "substantial utility" for his or her invention, i.e. a utility "where specific benefit exists in currently available form."[2] The Court

[1] *Brenner v. Manson,* 383 U.S. 519, 148 U.S.P.Q. (BNA) 689 (1966).

[2] *Id.* at 534, 148 U.S.P.Q. (BNA) at 695.

concluded that "a patent is not a hunting license. It is not a reward for the search, but compensation for its successful conclusion. A patent system must be related to the world of commerce rather than the realm of philosophy."[3]

Later, in *Nelson v. Bowler,*[4] the C.C.P.A. acknowledged that tests evidencing pharmacological activity of a compound may establish practical utility, even though they may not establish a specific therapeutic use. The court held that "since it is crucial to provide researchers with an incentive to disclose pharmaceutical activities in as many compounds as possible, we conclude adequate proof of any such activity constitutes a showing of practical utility."[5]

In *Cross v. Iizuka,*[6] the C.A.F.C. reaffirmed *Nelson*, and added that *in vitro* results might be sufficient to support practical utility, explaining that "*in vitro* testing, in general, is relatively less complex, less time consuming, and less expensive than *in vivo* testing. Moreover, *in vitro* results with the particular pharmacological activity are generally predictive of *in vivo* test results, i.e. there is a reasonable correlation there between."[7] The court perceived "No insurmountable difficulty" in finding that, under appropriate circumstances, "in vitro testing, may establish a practical utility."[8]

The case law has also clearly established that Appellants' statements of utility are usually sufficient, unless such statement of utility is unbelievable on its face.[9] The PTO has the initial burden to prove that Appellants' claims of usefulness are not believable on their face.[10] In general, an Appellant's assertion of utility creates a presumption of utility that will be sufficient

---

[3] *Id.* at 536, 148 U.S.P.Q. (BNA) at 696.

[4] *Nelson v. Bowler*, 626 F.2d 853, 206 U.S.P.Q. (BNA) 881 (C.C.P.A. 1980).

[5] *Id.* at 856, 206 U.S.P.Q. (BNA) at 883.

[6] *Cross v. Iizuka*, 753 F.2d 1047, 224 U.S.P.Q. (BNA) 739 (Fed. Cir. 1985).

[7] *Id.* at 1050, 224 U.S.P.Q. (BNA) at 747.

[8] *Id.*

[9] *In re Gazave*, 379 F.2d 973, 154 U.S.P.Q. (BNA) 92 (C.C.P.A. 1967).

[10] *Ibid.*

to satisfy the utility requirement of 35 U.S.C. §101, "unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope."[11,12]

Compliance with 35 U.S.C. §101 is a question of fact.[13] The evidentiary standard to be used throughout *ex parte* examination in setting forth a rejection is a preponderance of the totality of the evidence under consideration.[14] Thus, to overcome the presumption of truth that an assertion of utility by the Appellant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Only after the Examiner made a proper *prima facie* showing of lack of utility, does the burden of rebuttal shift to the Appellant. The issue will then be decided on the totality of evidence.

The well established case law is clearly reflected in the Utility Examination Guidelines ("Utility Guidelines")[15], which acknowledge that an invention complies with the utility requirement of 35 U.S.C. §101, if it has at least one asserted "specific, substantial, and credible utility" or a "well-established utility." Under the Utility Guidelines, a utility is "specific" when it is particular to the subject matter claimed. For example, it is generally not enough to state that a nucleic acid is useful as a diagnostic without also identifying the conditions that are to be diagnosed.

In explaining the "substantial utility" standard, M.P.E.P. §2107.01 cautions, however, that Office personnel must be careful not to interpret the phrase "immediate benefit to the public" or similar formulations used in certain court decisions to mean that products or services based on the claimed invention must be "currently available" to the public in order to satisfy the utility requirement. "Rather, any reasonable use that an Appellant has identified for the invention that can be viewed as providing a public benefit should be accepted as sufficient, at least with regard

---

[11] *In re Langer*, 503 F.2d 1380,1391, 183 U.S.P.Q. (BNA) 288, 297 (C.C.P.A. 1974).

[12] *See also In re Jolles*, 628 F.2d 1322, 206 U.S.P.Q. 885 (C.C.P.A. 1980); *In re Irons*, 340 F.2d 974, 144 U.S.P.Q. 351 (1965); *In re Sichert*, 566 F.2d 1154, 1159, 196 U.S.P.Q. 209, 212-13 (C.C.P.A. 1977).

[13] *Raytheon v. Roper*, 724 F.2d 951, 956, 220 U.S.P.Q. (BNA) 592, 596 (Fed. Cir. 1983) cert. denied, 469 US 835 (1984).

[14] *In re Oetiker*, 977 F.2d 1443, 1445, 24 U.S.P.Q.2d (BNA) 1443, 1444 (Fed. Cir. 1992).

[15] 66 Fed. Reg. 1092 (2001).

to defining a 'substantial' utility."[16] Indeed, the Guidelines for Examination of Applications for Compliance With the Utility Requirement,[17] gives the following instruction to patent examiners: "If the Appellant has asserted that the claimed invention is useful for any particular practical purpose . . . and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility."

B. The Data and Documentary Evidence Supporting a Patentable Utility

Appellants respectfully submit that Appellants rely on the gene amplification data for patentable utility of the claimed antibodies that bind the PRO351 polypeptide, and that the gene amplification data for the gene encoding the PRO351 polypeptide is clearly disclosed in the instant specification under Example 114.

It was well known in the art at the time the invention was made that gene amplification is an essential mechanism for oncogene activation. The gene amplification assay is well-described in Example 114 of the present application. Example 114 discloses that the inventors isolated genomic DNA from a variety of primary cancers and cancer cell lines that are listed in Table 9, including primary lung and colon tumors of the type and stage indicated in Table 8. As a negative control, DNA was isolated from the cells of ten normal healthy individuals, which was pooled and used as a control. Gene amplification was monitored using real-time quantitative TaqMan™ PCR. Table 9 shows the resulting gene amplification data. Further, Example 114 explains that the results of TaqMan™ PCR are reported in ΔCt units, wherein one unit corresponds to one PCR cycle or approximately a 2-fold amplification relative to control, two units correspond to 4-fold amplification, 3 units to 8-fold amplification etc.

Appellants respectfully submit that a ΔCt value of at least 1.0, which is a **more than 2-fold increase**, was observed for PRO351 in at least ten of the lung tumors listed in Table 9. Table 9 teaches that the nucleic acids encoding PRO351 showed 1.02 to 1.46 ΔCt units which corresponds to $2^{1.02}$ to $2^{1.46}$ - fold amplification or 2.03 to 2.75 amplification in ten types of human primary lung tumors, LT9, LT10, LT11, LT13, LT15, LT16, LT17, LT18, LT19 and

---

[16] M.P.E.P. §2107.01.

[17] M.P.E.P. §2107 II(B)(1).

LT21. Accordingly, the present specification clearly discloses strong evidence that the gene encoding the PRO351 polypeptide is significantly amplified in a significant number of lung tumors.

It is also well known that gene amplification occurs in most solid tumors, and generally is associated with poor prognosis.

In support, Appellants have submitted, in their Response filed May 20, 2005, a Declaration by Dr. Audrey Goddard. Appellants particularly draw the Board's attention to page 3 of the Goddard Declaration which clearly states that:

> It is further my considered scientific opinion that an at least **2-fold increase** in gene copy number in a tumor tissue sample relative to a normal (*i.e.*, non-tumor) sample **is significant** and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. Accordingly, a gene identified as being amplified at least 2-fold by the quantitative TaqMan PCR assay in a tumor sample relative to a normal sample is **useful as a marker for the diagnosis of cancer**, for monitoring cancer development and/or for measuring the efficacy of cancer therapy. (Emphasis added).

As indicated above, the gene encoding the PRO351 polypeptide shows significantly higher than a two fold amplification in ten different lung tumors. In addition, the Goddard Declaration clearly establishes that the TaqMan real-time PCR method described in Example 114 has gained wide recognition for its versatility, sensitivity and accuracy, and is in extensive use for the study of gene amplification. The facts disclosed in the Declaration also confirm that based upon the gene amplification results, one of ordinary skill would find it credible that PRO351 is a diagnostic marker of lung cancer.

The Examiner has asserted that "the specification provides data showing a very small increase in DNA copy number, approximately 2-fold, in a few tumor samples for PRO351." (Page 2 of the Advisory Action mailed October 14, 2004). The Examiner further asserts that "it was imperative to find evidence in the relevant scientific literature whether or not a small increase in DNA copy number would be considered by the skilled artisan to be predictive of

increased mRNA and polypeptide levels." (Page 3 of the Advisory Action mailed October 14, 2004).

Appellants respectfully submit that the Examiner seems to be applying a heightened utility standard in this instance, which is legally incorrect. Appellants have shown that the gene encoding PRO351 demonstrated significant amplification, from 2.03 to 2.75 fold, in three lung tumors. As explained in the Declaration of Dr. Audrey Goddard (submitted with the Response filed May 20, 2005):

> It is further my considered scientific opinion that an at least **2-fold increase** in gene copy number in a tumor tissue sample relative to a normal (*i.e.*, non-tumor) sample **is significant** and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. (Emphasis added).

By referring to the 2.03-fold to 2.75-fold amplification of the PRO351 gene in lung tumors as "very small" the Examiner appears to ignore the teachings within an expert's declaration without any basis, or without presenting any evidence to the contrary. Appellants respectfully draw the Board's attention to the Utility Examination Guidelines (Part IIB, 66 Fed. Reg. 1098 (2001)) which state that:

> Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered.

Thus, barring evidence to the contrary, Appellants maintain that the 2.03 to 2.75-fold amplification disclosed for the PRO351 gene is significant and forms the basis for the utility claimed herein.

1. The pooled normal blood sample is a valid negative control for gene amplification experiments

The Examiner has noted that "Table 9 reports a comparison of lung tumor tissue samples with a pooled sample of DNA from normal cells but not matched tissue samples (i.e., normal lung epithelium tissue)" and has asserted that "it is not clear if Dr. Goddard intended the phrase 'normal samples' to include unrelated tissue samples such as those used in the specification."

The Examiner concluded that "the art does not consider pooled, unrelated DNA samples to be an appropriate control." (Page 5 of the Office Action mailed August 29, 2005).

Appellants respectfully submit that the negative control taught in the specification was known in the art at the time of filing, and accepted as a true negative control as demonstrated by use in peer reviewed publications. For example, Pennica *et al.*, made of reference by the Examiner in the Office Action mailed February 24, 2004, explain that "[t]he relative WISP gene copy number in each colon tumor DNA was compared with **pooled normal DNA** from 10 donors by quantitative PCR" (page 14720, col. 2; emphasis added). Pennica *et al.* further explain that DNA was isolated from "the pooled blood of 10 normal human donors" (page 14718, col. 1). Thus Pennica *et al.* used the **same** control for their gene amplification experiments as that described in the instant specification.

In further examples, Pitti *et al.* (Exhibit F submitted with the Response filed May 20, 2005), used the same quantitative TaqMan PCR assay described in the specification to study gene amplification in lung and colon cancer of DcR3, a decoy receptor for Fas ligand. As described, Pitti *et al.* analyzed DNA copy number 'in genomic DNA from 35 primary lung and colon tumours, relative to pooled genomic DNA from peripheral blood leukocytes (PBL) of 10 healthy donors." (Page 701, col. 1; emphasis added). The authors also analyzed mRNA expression of DcR3 in primary tumor tissue sections and found tumor-specific expression, confirming the finding of frequent amplification in tumors, and confirming that the pooled blood sample was a valid negative control for the gene amplification experiments. In Bieche *et al.* (Exhibit G submitted with the Response filed May 20, 2005), the authors used the quantitative TaqMan PCR assay to study gene amplification of *myc*, *ccnd1* and *erbB2* in breast tumors. As their negative control, Bieche *et al.* used normal leukocyte DNA derived from a small subset of the breast cancer patients (page 663). The authors note that "[t]he results of this study are consistent with those reported in the literature" (page 664, col. 2), thus confirming the validity of the negative control. Accordingly, the art demonstrates that pooled normal blood samples are considered to be a valid negative control for gene amplification experiments of the type described in the specification.

The Examiner has asserted that "both Pitti *et al.* and Bieche *et al.* did not rely solely upon the PCR assay using a control from blood genomic DNA to make conclusions." The Examiner notes in particular that "Pitti *et al.* also looked at northern blot analysis, ligand binding analysis, apoptosis induction analysis, and *in situ* hybridization analysis" (Page 2 of the Advisory Action mailed December 7, 2005).

Appellants respectfully point out that northern blot analysis and *in situ* hybridization analysis were used to measure RNA levels, while ligand binding analysis and apoptosis induction analysis were used to measure the activity of the encoded protein. The sole technique that Pitti *et al.* relied upon to measure gene amplification was the PCR assay using the pooled blood control. While the Examiner notes that the authors used an additional control in the PCR assays, using flanking DNA regions in tumor samples compared to blood DNA samples, Appellants respectfully point out that because this assay also used the pooled normal blood sample as the comparison, it only supports Appellants' point that pooled normal blood sample is a valid control.

The Examiner further asserted that "Bieche *et al.* relied upon Southern blotting to confirm the PCR results and note that not all samples showing PCR amplification also showed amplification by Southern blotting," adding that "[t]his was especially true for sequences that were amplified at low levels comparable to the levels that instant PRO351 was shown to be amplified." (Page 3 of the Advisory Action mailed December 7, 2005). Appellants respectfully direct the Board's attention to the paragraph immediately following that referenced by the Examiner, where the authors explain that "[g]ene amplification status has been studied mainly by means of Southern blotting, but **this method is not sensitive enough to detect low-level gene amplification**, nor accurate enough to quantify the full range of amplification values" (page 664, col. 1; emphasis added). Thus the fact that the samples showing lower levels of gene amplification by PCR analysis did not necessarily also show amplification by Southern blotting is precisely what would be expected given that the PCR analysis is a **more sensitive technique**, and can reliably detect levels of amplification missed by the less sensitive Southern blotting technique. The fact that, as Bieche *et al.* report, "[t]he results of this study are consistent with

those reported in the literature" (page 664, col. 2), confirms the validity of the pooled normal blood control used in the PCR experiments.

Finally, the Examiner has noted that "publications have been cited as evidence that matched, cancer-free tissue samples are used as controls." (Page 3 of the Advisory Action mailed December 7, 2005). Appellants do not dispute that such matched, cancer-free samples may be used as controls. Appellants argue that pooled normal blood samples may also be used as an equally valid control, and have cited publications as evidence that such pooled normal blood samples are in fact used in the art as controls.

2. Aneuploidy is associated with cancerous or precancerous tissues, not normal tissues

The Examiner has asserted that "damaged, precancerous lung epithelium is often aneuploid," and stated that "[o]ne skilled in the art would not conclude that PRO351 is a diagnostic probe for lung cancer unless it is clear that PRO351 is amplified to a clearly greater extent in true lung tumor tissue relative to non-cancerous lung epithelium." (Page 4 of the Office Action mailed August 29, 2005). In support of this assertion the Examiner cited a reference by Hittelman.

Appellants note that the title of the Hittelman paper is "Genetic Instabilities in Epithelial Tissues at Risk for Cancer." Hittelman studied lung tissue from chronic smokers, which had been exposed for years to carcinogenic tobacco smoke. As Hittelman explains, "[t]umors of the aerodigestive tract have been proposed to reflect a 'field cancerization' process whereby the whole tissue is exposed to carcinogenic insult (e.g., tobacco smoke) and is at increased risk for multistep tumor development" (page 3). The detection of increases in chromosome number therefore identifies cells which have begun the first steps in this multistep progression to cancer. Even if these particular epithelial regions are not yet cancerous, their presence is strongly correlated with the development of cancer in the target tissue as a whole. Accordingly, Hittelman concludes that "**the measurement of chromosome instability in the target tissue will be useful in assessing cancer risk** as well as response to intervention" (page 10; emphasis added).

Accordingly, Hittelman shows that an increase in chromosome number or gene amplification is associated not with normal tissues, but with cancerous, or pre-cancerous tissues, and therefore, an increase in chromosome number or gene amplification is a useful marker for a cancerous or pre-cancerous state. Detection of pre-cancerous cells or tissues is useful because, as explained by Hittelman, it allows for assessing cancer risk, as well as response to intervention. Hence, Appellants respectfully submit that whether a pre-cancerous or tumor sample were analyzed, the showing of DNA amplification of the PRO351 gene would still be significant, since it would lead to the diagnosis of either a pre-cancerous state or a cancerous state, which is the utility asserted here. Despite the Examiner's assertion that such a use "is not well-established in the prior art," (page 4 of the Office Action mailed August 29, 2005) it is clear, as discussed above, that the use of amplified genes as markers for assessing cancer risk is explicitly contemplated in Hittelman *et al.* Thus even if the DNA amplification observed for PRO351 was correlated to pre-cancerous rather than cancerous lung tissue, it would still provide utility for PRO351.

C. A *prima facie* case of lack of utility has not been established

The Examiner has asserted that the disclosed gene amplification data does not establish a patentable utility for the PRO351 polypeptide or the claimed antibodies that bind it because allegedly "[t]he art demonstrates that gene amplification does not reliably correlate with increased mRNA transcript levels or increased polypeptide levels," citing references by Pennica *et al.* and Konopka *et al.* (Page 4 of the Office Action mailed February 24, 2004). Haynes *et al.* was cited "as providing evidence that polypeptide expression levels cannot be accurately predicted from mRNA levels" (Page 3 of the Advisory Action mailed October 14, 2004). In further support of the alleged lack of correlation between mRNA levels and polypeptide levels, the Examiner has cited additional references by Hu *et al.*, LaBaer, Chen *et al*, Lian *et al.*, Fessler *et al.*, Gygi *et al.* and Greenbaum *et al.*

As a preliminary matter, Appellants respectfully submit that it is not a legal requirement to establish that gene amplification necessarily results in increased expression at the mRNA and polypeptide levels, or that protein levels can be "accurately predicted." As discussed above, the evidentiary standard to be used throughout *ex parte* examination of a patent application is a

preponderance of the totality of the evidence under consideration. Accordingly, Appellants submit that in order to overcome the presumption of truth that an assertion of utility by the Appellant enjoys, the Examiner must establish that **it is more likely than not** that one of ordinary skill in the art would doubt the truth of the statement of utility. Therefore, it is not legally required that there be a "necessary" correlation between the data presented and the claimed subject matter. The law requires only that one skilled in the art should accept that such a correlation is **more likely than not to exist**. Appellants respectfully submit that when the proper evidentiary standard is applied, a correlation must be acknowledged.

1. *Pennica* et al. *and Konopka* et al.

In support of the assertion that "[t]he art demonstrates that gene amplification does not reliably correlate with increased mRNA transcript levels or increased polypeptide levels," the Examiner has cited references by Pennica *et al.* and Konopka *et al.* (Page 4 of the Office Action mailed February 24, 2004). In particular, the Examiner cited the abstract of Pennica *et al.* for its disclosure that "WISP-1 gene amplification and overexpression in human colon tumors showed a correlation between DNA amplification and over-expression, whereas overexpression of WISP-3 RNA was seen in the absence of DNA amplification. In contrast, WISP-2 DNA was amplified in colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with expression in normal colonic mucosa from the same patient." From this, the Examiner correctly concluded that increased copy number does not *necessarily* result in increased polypeptide expression. The standard, however, is not absolute certainty.

In fact, as noted even in Pennica *et al.*, "[a]n analysis of *WISP*-1 gene amplification and expression in human colon tumors *showed a correlation between DNA amplification and over-expression...*" (Pennica *et al.*, page14722, left column, first full paragraph, emphasis added). Thus the findings of Pennica *et al.* with respect to WISP-1 support Appellants' arguments. In the case of WISP-3, the authors report that there was no change in the DNA copy number, but there was a change in mRNA levels. This apparent lack of correlation between DNA and mRNA levels is not contrary to Appellants' assertion that a change in DNA copy number generally leads to a change in mRNA level. Appellants are not attempting to predict the DNA copy number based on changes in mRNA level, and Appellants have not asserted that the only means for

changing the level of mRNA is to change the DNA copy number. Therefore a change in mRNA without a change in DNA copy number is not contrary to Appellants' assertions.

The fact that the single WISP-2 gene did not show the expected correlation of gene amplification with the level of mRNA/protein expression does not establish that it is more likely than not, in general, that such correlation does not exist. The Examiner has not shown whether the lack or correlation observed for the WISP-2 gene is typical, or is merely a discrepancy, an exception to the rule of correlation. Indeed, the working hypothesis among those skilled in the art is that, if a gene is amplified in cancer, the encoded protein is likely to be expressed at an elevated level, as was demonstrated for WISP-1.

Accordingly, Appellants respectfully submit that Pennica *et al.* teaches nothing conclusive regarding the absence of correlation between amplification of a gene and over-expression of the encoded WISP polypeptide. More importantly, the teaching of Pennica *et al.* is specific to *WISP* genes. Pennica *et al.* has no teaching whatsoever about the correlation of gene amplification and protein expression in general.

The Examiner argues that Pennica *et al* is relevant even though it is limited to only one gene family because it is "evidence that one skilled in the art cannot assume that any one gene's amplification results in protein over-expression" and because the instant case also concerns a single gene. (Page 5 of the Office Action mailed August 29, 2005) Appellants disagree. The test is whether it is more likely than not that gene amplification results in overexpression of the corresponding mRNA and protein. In order to meet that standard, the Examiner must provide evidence that it is more likely than not that gene amplification does not result in mRNA or protein overexpression. Providing the single example of the WISP-2 gene does not suffice to meet this burden.

Appellants next respectfully submit that, contrary to the PTO's assertions, Konopka *et al.* **supports** Appellants' position that mRNA levels correlate with protein levels. Konopka *et al.* states that "the 8-kb mRNA that encodes $P210^{c\text{-}abl}$ was detected at a 10-fold higher level in SK-CML7bt-333 (Fig. 3A, +) than in SK-CML16Bt-1 (B, +), which **correlated** with the relative level of $P210^{c\text{-}abl}$ detected in each cell line. Analysis of additional cell lines demonstrated that

the level of 8-kb mRNA **directly correlated** with the level of $P210^{c\text{-}abl}$ (Table 1)" (page 4050, col. 2, emphasis added).

Nor does Konopka *et al.* support the PTO's position that DNA amplification is not correlated with mRNA or protein overexpression. Konopka *et al.* show only that, of the cell lines known to have increased abl protein expression, only one had amplification of the abl gene (page 4051, col. 1). This result proves only that increased mRNA and protein expression levels can result from causes other than gene amplification. Konopka *et al.* do not demonstrate that when gene amplification does occur, it does not result in increased mRNA and protein expression levels, particularly given that the cell line with amplification of the abl gene did show increased abl mRNA and protein expression levels.

2. *Hu* et al. *and LaBaer*

The Examiner refers to the reference by Hu et al. as allegedly demonstrating that it is incorrect that increased mRNA production leads to increased protein production. In particular, the Examiner cites Hu *et al.* to the effect that genes displaying a 5-fold change or less in mRNA expression in tumors compared to normal showed no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease. (Page 9 of the Office Action mailed August 19, 2005).

Appellants submit that in order to overcome the presumption of truth that an assertion of utility by the Appellant enjoys, the Examiner must establish that it is more likely than not that one of ordinary skill in the art would doubt the truth of the statement of utility. Accordingly, contrary to the Examiner's assertion, Appellants submit that Hu *et al.* does not show a lack of correlation between microarray data and the biological significance of cancer genes.

Appellants respectfully point out that the analysis by Hu *et al.* has certain statistical flaws. According to Hu *et al.*, "*different* statistical methods 'were applied to' *estimate* the strength of gene-disease relationships and evaluated the results." (See page 406, left column, emphasis added). Using these different statistical methods, Hu *et al.* "[a]ssessed the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation." (See page 411, left column). It is well known in the art that various statistical methods allow different

variables to be manipulated to affect the outcome. For example, the authors admit, "Initial attempts to search the literature using" the list of genes, gene names, gene symbols, and frequently used synonyms, generated by the authors "revealed several sources of false positives and false negatives." (See page 406, right column). The authors further admit that the false positives caused by "duplicative and unrelated meanings for the term" were "difficult to manage." Therefore, in order to minimize such false positives, Hu *et al.* disclose that these terms "had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes." (See page 406, right column). Hence, Appellants respectfully submit that in order to minimize the false positives and negatives in their analysis, Hu *et al.* manipulated various aspects of the input data.

Appellants further submit that the statistical analysis by Hu *et al.* is not a reliable standard because the frequency of citation only reflects the current research interest in a molecule, not the true biological function of the molecule. Indeed, the authors acknowledge that "[r]elationships established by frequency of co-citation do not necessarily represent a true biological link." (See page 411, right column). One would expect that genes with the greatest change in expression in a disease would be the first targets of research, and therefore have the strongest known relationship to the disease as measured by the number of publications reporting a connection with the disease. The correlation reported in Hu only indicates that the greater the change in expression level, the more likely it is that there is a published or known role for the gene in the disease, as found by their automated literature-mining software. Thus, Hu's results merely reflect a bias in the literature toward studying the most prominent targets, and say nothing regarding the ability of a gene that is 2-fold or more differentially expressed in tumors to serve as a disease marker.

Even assuming that Hu *et al.* provide evidence to support a true relationship, the conclusion in Hu *et al.* only applies to a specific type of breast tumor (estrogen receptor (ER)-positive breast tumor) and can not be generalized as a principle governing microarray study of breast cancer in general, let alone the various other types of cancer genes in general. In fact, even Hu *et al* admit that., "[i]t is likely that this threshold will change depending on the disease as well as the experiment. Interestingly, the observed correlation was only found among

ER-positive (breast) tumors not ER-negative tumors." (See page 412, left column). Therefore, based on these findings, the authors add, "[t]his may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently." (See page 412, left column; emphasis added).

Furthermore, Hu *et al.* did not look for a correlation between changes in mRNA and changes in protein levels, and therefore their results are not contrary to Appellants' assertion that there is a correlation between the two. Appellants are not relying on any "biological role" that the PRO351 polypeptide has in cancer for its asserted utility. Instead, Appellants are relying on the overexpression of PRO351 in certain tumors compared to their normal tissue counterparts. Nowhere in Hu does it say that a lack of correlation in their study means that genes with a less than five-fold change in level of expression in cancer cannot serve as a diagnostic marker of cancer.

The Examiner asserts that "Appellant is holding Hu *et al.* to a higher standard than their own specification" for statistical analysis. (Page 10 of the Office Action mailed August 29, 2005). However, Appellants have compared the level of amplification of the PRO351 gene in normal and lung tumors and have provided information indicating a greater than 2 fold amplification. Appellants are not relying on statistical analysis of information obtained from published literature based on the current research interest of a molecule, and hence the issues regarding statistical analysis of such information do not apply to Appellants' data.

In the Advisory Action mailed December 7, 2005, the Examiner cited an additional article by LaBaer. Appellants respectfully point out that LaBaer is an unreviewed letter to the editor by an author of the Hu *et al.* article describing the automated literature searching tool used in the Hu *et al.* reference discussed above. LaBaer provides no further evidence than that provided in Hu, and provides no evidence whatsoever to support the conclusion that the results of Hu are applicable to the diagnostic utility of differentially expressed genes. Importantly, like the Hu reference, LaBaer does not consider or offer any discussion of whether there is a correlation between changes in mRNA levels and changes in the level of the encoded protein. In addition, LaBaer's conclusions regarding disease-independent differences between samples are not

applicable in the instant case where normal human tissue and human tumor tissue samples were compared. Accordingly, LaBaer suffers from the same defects discussed above with respect to Hu *et al.*

3. ***Chen* et al.**

In further support of the alleged lack of correlation between increased mRNA levels in cancer as compared to normal tissues and increased polypeptide levels, the Examiner cited Chen *et al.* as allegedly disclosing that "only 17% of 165 polypeptide spots or 21% of the genes had a significant correlation between protein and mRNA expression levels in lung adenocarcinoma samples." (Page 6 of the Advisory Action mailed December 7, 2005).

First, Appellants note that proteins selected for study by Chen *et al.* were those detectable by staining of 2D gels. As noted in, for example, Haynes *et al.*, made of record by the Examiner in the Advisory Action mailed December 7, 2005, there are problems with selecting proteins detectable by 2D gels. "It is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions." (page 1870, col. 1). Thus, Chen *et al.*, by selecting proteins detectable by staining of 2D gels, are likely to have excluded from their analysis many of the proteins most likely to be significant as cancer markers.

Secondly, Chen *et al.* looked at expression levels across a set of samples including a large number of tumor samples (76) along with a much smaller number of normal samples (9). The tumor samples were taken from stage 1 and stage III lung adenocarcinomas, which were classified as bronchoaveolar, bronchial derived or both bronchial and bronchoaveolar derived. Accordingly, the tissues examined were from different tissues in different stages of normal or cancerous growth. The authors determined the relationship between mRNA and protein expression by using the average expression values for all samples. The average value for each protein or mRNA was generated using all 85 lung tissue samples. This resulted in negative normalized protein values in some cases. Further, the authors chose an arbitrary threshold of 0.115 for the correlation to be considered significant. Accordingly, the Chen paper does not account for different expression in different tissues or different stages of cancer.

Thirdly, no attempt was made to compare expression levels in normal versus tumor samples, and in fact the authors concede that they had too few normal samples for meaningful analysis (page 310, col. 2). As a result, the analysis in the Chen paper shows only that a number of randomly selected proteins have varying degrees of correlation between mRNA and protein expression levels within a set of different lung adenocarcinoma samples. The Chen paper does not address the issue of whether increased mRNA levels in the tumor samples taken together as one group, as compared to the normal samples as a group, correlated with increased protein levels in tumorous versus normal tissue. Accordingly, the results presented in the Chen paper are not applicable to the application at issue.

The correct test of utility is whether the utility is "more likely than not". In the case of the Chen reference, even if the analysis presented is correct (which is disputed), a review of the correlation coefficient data presented in the Chen *et al.* paper indicates that it is more likely than not that increased mRNA expression correlates with increased protein expression. A review of Table 1, which lists 66 genes [the paper incorrectly states there are 69 genes listed] for which only one protein isoform is expressed, shows that 40 genes out of 66 had a positive correlation between mRNA expression and protein expression. This clearly meets the test of "more likely than not". Similarly, in Table II , 30 genes with multiple isoforms [again the paper incorrectly states there are 29] were presented. In this case, for 22 genes out of 30, at least one isoform showed a positive correlation between mRNA expression and protein expression. Furthermore, 12 genes out of 29 showed a strong positive correlation [as determined by the authors] for at least one isoform. No genes showed a significant negative correlation. It is not surprising that not all isoforms are positively correlated with mRNA expression. Certain isoforms are likely non-functional proteins. Thus, Table II also provides that it is more likely than not that protein levels will correlate with mRNA expression levels.

4. ***Haynes* et al. *and Gygi* et al.**

The Examiner has cited the Haynes reference to establish that "even if gene amplification correlates with increased transcription, it does not *always* follow that protein levels are also amplified." (Page 5 of the Office Action mailed February 24, 2004; emphasis added). The Examiner further states that "Haynes *et al.* was cited as providing evidence that polypeptide

levels cannot be accurately predicted from mRNA levels." (Page 3 of the Advisory Action mailed October 14, 2004).

Appellants respectfully point out that Haynes *et al.* never indicate that the correlation between mRNA and protein levels does not exist. Haynes *et al.* only state that "protein levels cannot be *accurately* predicted from the level of the corresponding mRNA transcript" (See page 1863, under Section 2.1, last line, emphasis added). This result is expected, since there are many factors that determine translation efficiency for a given transcript, or the half-life of the encoded protein. Not surprisingly, Haynes *et al.* concluded that protein levels cannot always be accurately predicted from the level of the corresponding mRNA transcript in a single cellular stage or type when looking at the level of transcripts across different genes.

Importantly, Haynes *et al.* did not say that for a single gene, a change in the level of mRNA transcript is not positively correlated with a change in the level of protein expression. Appellants have asserted that increasing the level of mRNA for a particular gene leads to a corresponding increase for the encoded protein. Haynes *et al.* did not study this issue and says absolutely nothing about it. One cannot look at the level of mRNA across several different genes to investigate whether a change in the level of mRNA for a particular gene leads to a change in the level of protein for that gene. Therefore, Haynes *et al.* is not inconsistent with or contradictory to the utility of the instant claims, and offers no support for the PTO's rejection of Appellants' asserted utility.

Furthermore, Appellants note that contrary to the Examiner's statement, Haynes teaches that "*there was a general trend* but no strong correlation between protein [expression] and transcript levels" (See page 1863, under Section 2.1, emphasis added). For example, in Figure 1, there is a positive correlation between mRNA and protein amongst *most* of the 80 yeast proteins studied but the correlation is not linear, hence the authors suggest that one cannot *accurately* predict protein levels from mRNA levels. In fact, very few data points deviated or scattered away from the expected normal or showed a lack of correlation between mRNA: protein levels. Thus, the Haynes data meets the "more likely than not standard" and shows that a positive correlation exists between mRNA and protein. Therefore, Appellants submit that the Examiner's rejection is based on a misrepresentation of the scientific data presented in Haynes *et al.*

Haynes *et al.* may teach that protein levels cannot be "accurately predicted" from mRNA levels in the sense that the exact numerical amounts of protein present in a tissue cannot be determined based upon mRNA levels. Appellants respectfully submit that the PTO's emphasis on the need to "accurately predict" protein levels based on mRNA levels misses the point. The asserted utility for the claimed polypeptides is in the diagnosis of cancer. What is relevant to use as a cancer diagnostic is relative levels of gene or protein expression, not absolute values, that is, that the gene or protein is differentially expressed in tumors as compared to normal tissues. Appellants need only show that there is a correlation between mRNA and protein levels, such that mRNA overexpression generally predict protein overexpression. A showing that mRNA levels can be used to "accurately predict" the precise levels of protein expression is not required.

In the Advisory Action mailed December 7, 2005, the Examiner cited Gygi *et al.*, a study on which the Haynes references is based. Like Haynes, the Gygi reference looked at levels of mRNA at the same growth phase across different genes, not changes in mRNA levels for a single gene. Thus, when Gygi *et al.* state that "the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data," the authors are referring to correlations between constant levels of mRNA and protein at the same growth phase across different genes, not a correlation between a change in mRNA level and a change in protein level for the same gene and corresponding protein. Therefore, for the same reasons that Haynes is not relevant to Appellants' asserted utility, Gygi likewise offers no support for the PTO's rejection of Appellants' asserted utility.

Furthermore, Appellants submit that Gygi *et al* too did not indicate that a correlation between mRNA and protein levels does not exist. Gygi *et al.* only state that the correlation may not be sufficient in **accurately** predicting protein level from the level of the corresponding mRNA transcript (Emphasis added) (see page 1270, Abstract). *Accurate prediction* is not a criteria that is necessary for meeting the utility standards. Appellants note that the Gygi data indicate **a general trend** of correlation between protein [expression] and transcript levels (Emphasis added). For example, as shown in Figure 5, an mRNA abundance of **250-300** copies /cell correlates with a protein abundance of **500-1000** x 10^3 copies/cell. An mRNA abundance of **100-200** copies/cell correlates with a protein abundance of **250-500** x 10^3

copies/cell (emphasis added). Therefore, high levels of mRNA **generally** correlate with high levels of proteins. In fact, most data points in Figure 5 did not deviate or scatter away from the general trend of correlation. Thus, the Gygi data meets the "more likely than not standard" and shows that a positive correlation exists between mRNA and protein. Therefore, Appellants submit that the Examiner's rejection is based on a misrepresentation of the scientific data presented in Gygi *et al.*

5. Lian et al.

In further support of the alleged lack of correlation between mRNA expression and protein expression levels, the PTO has cited Lian *et al.* for the statement that there is a poor correlation between mRNA expression and protein abundance in mouse cells, and therefore it may be difficult to extrapolate directly from individual mRNA changes to corresponding ones in protein levels. (Page 6 of the Office Action mailed August 29, 2005).

In Lian *et al.*, the authors looked at the mRNA and protein levels of genes in a derived promyelocytic mouse cell-line during differentiation of the cells from a promyelocytic stage of development to mature neutrophils following treatment with retinoic acid. The level of mRNA expression was measured using 3'-end differential display (DD) and oligonucleotide chip array hybridization to examine the expression of genes at 0, 24, 48 and 72 hours after treatment with retinoic acid. Protein levels were qualitatively assessed at 0 and 72 hours after retinoic acid treatment following 2-dimensional gel electrophoresis.

Lian *et al.* report that they were able to identify 28 proteins which they considered differentially expressed (page 521). Of those 28, only 18 had corresponding gene expression information, and only 13 had measurable levels of mRNA expression (page 521, Table 6). The authors then compared the qualitative protein level from the 2-D electrophoresis gel to the corresponding mRNA level, and reported that only 4 genes of the 18 present in the database had expression levels which were consistent with protein levels (page 521, col. 1). The authors note that "[n]one of these was on the list of genes that were differentially expressed significantly (5-fold or greater change by array or 2-fold or greater change by DD)" (page 521; emphasis added). Based on these data, the authors conclude "[f]or protein levels based on estimated intensity of

Coomassie dye staining in 2DE, there was poor correlation between changes in mRNA levels and estimated protein levels" (page 522, col. 2).

The authors themselves admit that there are a number of problems with the data presented in this reference. At page 520 of this article, the authors explicitly express their concerns by stating that "[t]hese data must be considered with several caveats: membrane and other hydrophobic proteins and very basic proteins are not well displayed by the standard 2DE approach, and proteins presented at low level will be missed. In addition, to simplify MS analysis, we used a Coomassie dye stain rather than silver to visualize proteins, and this decreased the sensitivity of detection of minor proteins." (emphasis added). It is known in the art that Coomassie dye stain is a very insensitive method of measuring protein. This suggests that the authors relied on a very insensitive measurement of the proteins studied. The conclusions based on such measurements can hardly be accurate or generally applicable. In particular, the total number of proteins examined by Lian *et al.* was only 50 (page 520, col. 2), as compared to the approximately 7000 genes for which mRNA levels were measured (page 515, col. 1). Thus the conclusions are based on a very small and atypical set of proteins.

Appellants also emphasize that Appellants are asserting that a measurable change in mRNA level generally leads to a corresponding change in the level of protein expression, not that changes in protein level can be used to predict changes in mRNA level. As discussed above, Lian *et al.* did not take genes which showed significant mRNA changes and check the corresponding protein levels. Instead, the authors looked at a small and unrepresentative number of proteins, and checked the corresponding mRNA levels. Based on the authors' criteria, mRNA levels were significantly changed if they were at least 5-fold different when measured using a microchip array, or 2-fold different when using the more sensitive 3'-end differential display (DD). Of the 28 proteins listed in Table 6, only one has an mRNA level measured by microarray which is differentially expressed according to the authors (spot 7: melanoma X-actin, for which mRNA changed from 2539 to 341.3, and protein changed from 1 to 3). None of the other mRNAs listed in Table 6 show a significant change in expression level when using the criteria established by the authors for the less sensitive microarray technique.

There is also one gene in Table 6 whose expression was measured by the more sensitive technique of DD, and its level increased from a qualitative value of 0 to 2, a more than 2-fold increase (spot 2: actin, gamma, cytoplasmic). This increase in mRNA was accompanied by a corresponding increase in protein level, from 3 to 6.

Therefore, although the authors characterize the mRNA and protein levels as having a "poor correlation," this does not reflect a lack of a correlation between a change in mRNA level and a corresponding change in protein level. Only two genes meet the authors' criteria for differentially expressed mRNA level, and of those, one apparently shows a corresponding change in protein level and one does not. Thus, there is little basis for the authors' conclusion that "it may be difficult to extrapolate directly from individual mRNA changes to corresponding ones in protein levels (as estimated from 2DE)."

Finally, Appellants submit that Lian *et al.* only teach that protein expression may not correlate with mRNA level in differentiating myeloid cells and does not teach anything regarding such a lack of correlation for genes in general. Myeloid cell differentiation relates to hematopoiesis and is an entirely different biological process from solid tumor development because these two process involve entirely different regulatory mechanisms and molecules. Analysis of surface antigens expressed on myeloid cells of the granulocyte-monocyte-histiocyte series during differentiation in normal and malignant myelomonocytic cells is useful in identifying and classifying human leukemias and lymphomas, but cannot be used in diagnosis of any solid tumors. Therefore, even if the teaching of Lian *et al.* accurately reflects the correlation between mRNA and protein for the particular system studied, it can not apply to the tumor diagnosis assays of the present application.

6. Fessler et al.

In further support of the alleged lack of correlation between mRNA expression and protein expression levels, the PTO has also cited a publication by Fessler *et al.*, as having "found a 'poor concordance between mRNA transcript and protein expression changes' in human cells." (Page 6 of the Office Action mailed August 29, 2005). Fessler is not contrary to Appellants' asserted utility, and actually supports Appellants' assertion that a change in the level of mRNA for a particular protein generally leads to a corresponding change in the level of the encoded

protein. As noted above, Appellants make no assertions regarding changes in protein levels when mRNA levels are unchanged, nor does evidence of changes in protein levels when mRNA levels are unchanged have any relevance to Appellants' asserted utility.

Fessler *et al.* studied changes in neutrophil (PMN) gene transcription and protein expression following lipoplysaccharide (LPS) exposure. In Table VIII, Fessler *et al.* list a comparison of the change in the level of mRNA for 13 up-regulated proteins and 5 down-regulated proteins. Of the 13 up-regulated proteins, a change in mRNA levels is reported for only 3 such proteins. For these 3, mRNA levels are increased in 2 and decreased in the third. Of the 5 down-regulated proteins, a change in mRNA is reported for 3 such proteins. In all 3, mRNA levels also are decreased. Thus, in 5 of the 6 cases for which a change in mRNA levels are reported, the change in the level of mRNA corresponds to the change in the level of the protein. This is consistent with Appellants' assertion that a change in the level of mRNA for a particular protein generally leads to a corresponding change in the level of the encoded protein.

Regarding the remainder of the proteins listed in Table VIII, in 6 instances, protein levels changed while mRNA levels were unchanged. This evidence has no relevance to Appellants' assertion that changes in mRNA levels lead to corresponding changes in protein levels, since Appellants are not asserting that changes in mRNA levels are the only cause of changes in protein levels. In the final 6 instances listed in Table VIII, protein levels changed while mRNA was noted as "absent." This evidence also has no relevance to Appellants' assertion that changes in mRNA levels causes corresponding changes in protein levels. By virtue of being "absent," it is not possible to tell whether mRNA levels were increased, decreased or remained unchanged in PMN upon contact with LPS. Nothing in these results by Fessler *et al.* suggests that a change in the level of mRNA for a particular protein does not generally lead to a corresponding change in the level of the encoded protein. Accordingly, these results are not contrary to Appellants' assertions.

The PTO points to Fessler's statement regarding Table VIII that there was "a poor concordance between mRNA transcript and protein expression changes." (Page 6 of the Office Action mailed August 29, 2005). As is clear from the above discussion, this statement does not relate to a lack of correlation between a change in mRNA levels leading to a change in protein

levels, because in 5 of 6 such instances, changes in mRNA and protein levels correlated well. Instead, this statement relates to observations in which protein levels changed when mRNA was either unchanged or "absent." As such, this statement is an observation that in addition to transcriptional activity, LPS also has post-transcriptional and possibly post-translational activity that affect protein levels, an observation which is not contrary to Appellants' assertions. Accordingly, Fessler's results are consistent with Appellants' assertion that a change in mRNA level of for a particular protein generally leads to a corresponding change in the level of the encoded protein, since 5 of 6 genes demonstrated such a correlation.

7. ***Greenbaum* et al.**

In further support of the alleged lack of correlation between mRNA expression and protein expression levels, the Examiner cited new reference by Greenbaum *et al.* The Examiner asserted that Greenbaum *et al.* teaches that, "To date, there have been only a handful of efforts to find correlations between mRNA and protein expression levels. And, for the most part, they have reported only minimal and/or limited correlations." (Page 4 of the Advisory Action mailed December 7, 2005).

Appellants note that Greenbaum *et al.* compared the expression of a number of different mRNAs and their corresponding proteins in yeast cells. Greenbaum *et al.* did not compare the change of expression of specific mRNAs and their corresponding proteins in cancer cells versus normal cells. Accordingly, this reference is also not relevant to the issue at hand. Nevertheless, Greenbaum states that logically "we would assume that those ORFs that show a large degree of variation in their expression are controlled at the transcriptional level. The variability of the mRNA expression is indicative of the cell controlling the mRNA expression at different points of the cell cycle to achieve the resulting and desired protein. **Thus we would expect and we found a high degree of correlation (r-0.89) between the reference mRNA and protein levels for these particular ORFs: the cell has already put significant energy into dictating the final level of protein through tightly controlling the mRNA expression**" (page 117.5, col. 1; emphasis added). Furthermore, Greenbaum states that **"we found that ORFs that have higher than average levels of ribosomal occupancy – that is that a large percentage of their cellular mRNA concentration is associated with ribosomes (being translated) – have well correlated**

mRNA and protein expression levels. (Figure 2)." (page 117.5, col. 2; emphasis added). Therefore, contrary to the Examiner's assertion, Greenbaum does find high levels of correlation between mRNA and protein expression in yeast cells. In particular, Greenbaum demonstrates that a high degree of correlation is found for those genes which show a large degree of variability in mRNA expression – that is, for those genes which show changes in mRNA expression, the change in mRNA expression is correlated with a change in protein expression.

In summary, Appellants respectfully submit that the Examiner has not shown that gene amplification in tumor as compared to normal tissue is not correlated with changes in mRNA and protein expression. The Patent Office has failed to meet its initial burden of proof that Appellants' claims of utility are not substantial or credible. The arguments presented by the Examiner in combination with the Pennica, Konopka, Hu, LaBaer, Chen, Haynes, Gygi, Lian, Fessler, and Greenbaum articles do not provide sufficient reasons to doubt the statements by Appellants that PRO351 has utility. As discussed above, the law does not require the existence of a "necessary" correlation between mRNA and protein levels. Nor does the law require that protein levels be "accurately predicted." According to the authors themselves, the data in the above cited references confirm that there is a general trend between protein expression and transcript levels, which meets the "more likely than not standard" and show that a positive correlation exists between mRNA and protein. Therefore, Appellants submit that the Examiner's reasoning is based on a misrepresentation of the scientific data presented in the above cited reference and application of an improper, heightened legal standard. In fact, contrary to what the Examiner contends, the art indicates that, if a gene is overexpressed in cancer, it is more likely than not that the encoded protein will also be expressed at an elevated level.

D. It is "more likely than not" for amplified genes to have increased mRNA and protein levels

Appellants have submitted ample evidence to show that, in general, if a gene is amplified in cancer, it is more likely than not that the encoded protein will be expressed at an elevated level. First, the articles by Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.*, (made of record in Appellants' Response filed July 28, 2004) collectively teach that in general, gene amplification increases mRNA expression. Second, the Declaration of Dr. Paul Polakis, principal investigator

of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, shows that, in general, there is a correlation between mRNA levels and polypeptide levels. Thus, taken together, all of the submitted evidence supports Appellants' position that gene amplification is more likely than not predictive of increased mRNA and polypeptide levels.

Appellants submit that there are numerous articles which show that generally, if a gene is amplified in cancer, it is more likely than not that the mRNA transcript will be expressed at an elevated level. For example, Orntoft *et al.* (*Mol. and Cell. Proteomics*, 2002, vol. 1, pages 37-45 made of record in Appellants' Response filed July 28, 2004) studied transcript levels of 5600 genes in malignant bladder cancers, many of which were linked to the gain or loss of chromosomal material using an array-based method. Orntoft *et al.* showed that there was a gene dosage effect and taught that "in general (18 of 23 cases) chromosomal areas with more than 2-fold gain of DNA showed a corresponding increase in mRNA transcripts" (see column 1, abstract). In addition, Hyman *et al.* (*Cancer Res.*, 2002, vol. 62, pages 6240-45 - made of record in Appellants' Response filed July 28, 2004) showed, using CGH analysis and cDNA microarrays which compared DNA copy numbers and mRNA expression of over 12,000 genes in breast cancer tumors and cell lines, that there was "evidence of a prominent global influence of copy number changes on gene expression levels." (See page 6244, column 1, last paragraph). Additional supportive teachings were also provided by Pollack *et al.*, (*PNAS*, 2002, vol. 99, pages 12963-12968 - made of record in Appellants' Response filed July 28, 2004) who studied a series of primary human breast tumors and showed that "...62% of highly amplified genes show moderately or highly elevated expression, and DNA copy number influences gene expression across a wide range of DNA copy number alterations (deletion, low-, mid- and high-level amplification), and that on average, a 2-fold change in DNA copy number is associated with a corresponding 1.5-fold change in mRNA levels." Thus, these articles collectively teach that in general, gene amplification increases mRNA expression.

In addition, in their Response filed July 28, 2004, Appellants submitted a Declaration by Dr. Polakis, principal investigator of the Tumor Antigen Project of Genentech, Inc., the assignee of the present application, to show that mRNA expression correlates well with protein levels, in general. As Dr. Polakis explains, the primary focus of the microarray project was to identify

tumor cell markers useful as targets for both the diagnosis and treatment of cancer in humans. The scientists working on the project extensively rely on results of microarray experiments in their effort to identify such markers. As Dr. Polakis explains, using microarray analysis, Genentech scientists have identified approximately 200 gene transcripts (mRNAs) that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. To the date of the Declaration, they have generated antibodies that bind to about 30 of the tumor antigen proteins expressed from these differentially expressed gene transcripts and have used these antibodies to quantitatively determine the level of production of these tumor antigen proteins in both human cancer cells and corresponding normal cells. Having compared the levels of mRNA and protein in both the tumor and normal cells analyzed, they found a very good correlation between mRNA and corresponding protein levels. Specifically, in approximately 80% of their observations they have found that increases in the level of a particular mRNA correlates with changes in the level of protein expressed from that mRNA. While the proper legal standard is to show that the existence of correlation between mRNA and polypeptide levels is more likely than not, the showing of approximately 80% correlation for the molecules tested according to the Polakis Declaration greatly exceeds this legal standard. Based on these experimental data and his vast scientific experience of more than 20 years, Dr. Polakis states that, for human genes, increased mRNA levels typically correlate with an increase in abundance of the encoded protein. He further confirms that "it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded protein."

Appellants further note that the sale of gene expression chips to measure mRNA levels is a highly successful business, with a company such as Affymetrix recording 168.3 million dollars in sales of their GeneChip arrays in 2004. Clearly, the research community believes that the information obtained from these chips is useful (i.e., that it is more likely than not informative of the protein level).

In the Advisory Action mailed October 14, 2004, the Examiner asserted that "Orntoft *et al.* do not appear to look at gene amplification, mRNA levels and polypeptide levels from a single gene at a time.... Orntoft *et al.* concentrated on regions of chromosomes with strong gains of chromosomal material containing clusters of genes (p.40). This analysis was not done for

PRO351 in the instant specification. That is, it is not clear whether or not PRO351 is in a gene cluster in a region of a chromosome that is highly amplified. Therefore, the relevance of Orntoft *et al.* is not clear." (Pages 4-5 of the Advisory Action mailed October 14, 2004). The Examiner further asserted that "Hyman *et al.* used the same CGH approach in their research. Less than half (44%) of highly amplified genes showed mRNA overexpression (abstract).... Therefore, Hyman *et al.* also do not support utility of the polypeptides of the instant invention." (Page 5 of the Advisory Action mailed October 14, 2004). The Examiner further asserted that "Pollack *et al*, also used CGH technology, concentrating on large chromosome regions showing high amplification (p. 12965). Pollack *et al.* did not investigate polypeptide levels. Therefore, Pollack *et al.* also do not support the asserted utility of the claimed invention." (Page 5 of the Advisory Action mailed October 14, 2004).

In Orntoft *et al.*, 1,800 genes that yielded an increase or decrease in mRNA expression in two invasive tumors compared to the two non-invasive papillomas were then mapped to chromosomal locations. The chromosomes had already been analyzed for amplification by hybridizing tumor DNA to normal metaphase chromosomes (CGH). Orntoft *et al.* used CGH alterations as the independent variable and estimated the frequency of expression alterations of the 1,800 genes in the chromosomal areas. Orntoft *et al.* found that in general (77% and 80% concordance) areas with a strong gain of chromosomal material contained a cluster of genes having increased mRNA expression (see page 40). Orntoft *et al.* state, "For both tumors TCC733 ($p<0.015$) and TCC827 ($p<0.00003$) a highly significant correlation was observed between the level of CGH ratio change (reflecting the DNA copy number) and alterations detected by the array based technology" (see page 41, column 1). Orntoft *et al.* also studied the relationship between altered mRNA and protein levels using 2D-PAGE analysis. Orntoft *et al.* state, "In general there was a highly significant correlation ($p<0.005$) between mRNA and protein alterations.... 26 well focused proteins whose genes had a known chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ($p<0.005$) with the mRNA changes detected using the arrays." (See page 42, column 2 to page 34, column 2). Accordingly, Orntoft *et al.* clearly support Appellants' position that proteins expressed by genes that are amplified in tumors are useful as cancer markers.

The Examiner indicates that Appellants have not indicated whether PRO351 is in a gene cluster region of a chromosome. (Page 5 of the Advisory Action mailed October 14, 2004). Appellants fail to see how this is relevant to the analysis. Orntoft *et al.* did not limit their findings to only those regions of amplified gene clusters. Further, as discussed below, Hyman *et al.* and Pollack *et al.* did gene-by-gene analysis across all chromosomes.

The Examiner has asserted that "Orntoft *et al. could only compare the levels of about 40 well-resolved and focused abundant proteins.*" (Page 7 of the Office Action mailed August 29, 2005; emphasis in original). The Examiner further asserts that "Appellants have provided no fact or evidence concerning a lack of correlation between the specification's disclosure of low levels of amplification of DNA (which were not characterized on the basis of those in the Orntoft publication) and an associated rise in level of the encoded protein." (Page 7 of the Office Action mailed August 29, 2005).

Appellants respectfully point out that while technical considerations did prevent Orntoft *et al.* from evaluating a larger number of proteins, the ones they did look at showed a clear correlation between mRNA and protein expression levels. As Orntoft *et al.* state, "In general there was a highly significant correlation ($p<0.005$) between mRNA and protein alterations.... 26 well focused proteins whose genes had a known chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ($p<0.005$) with the mRNA changes detected using the arrays." (See page 42, column 2 to page 34, column 2). Accordingly, Orntoft *et al.* clearly support Appellants' position that proteins expressed by genes that are amplified in tumors are useful as cancer markers.

In addition, discussed above, the levels of amplification for PRO351 were **not** "low" but significant, and ranged from 2.03-fold to 2.75-fold. Appellants note that the levels of gene amplification observed by Orntoft *et al.* were relatively low, averaging only 0.3-0.4 fold (page 40, col. 1). In particular, the level of gene amplification associated with expression changes was only around two-fold (see Figure 2), as compared to the 2.04-fold to 2.75-fold amplification observed for PRO351. Even at these levels of gene amplification, Orntoft *et al.* found that "**[i]n most cases**, chromosomal gains detected by CGH were accompanied by an increased level of transcripts in both TCCs 733 (77%) and 827 (80%)" (page 40, col. 2; emphasis added). The

level of correlation between DNA copy number and increased mRNA levels observed by Orntoft *et al.*, from **77-80%**, clearly meets the standard of more likely than not. Orntoft *et al.* also found a "highly significant" correlation between mRNA and protein levels, with the two data sets studied having correlations of 39/40 (**98%**) and 19/26 (**73%**) (pages 42-43).

Appellants respectfully submit that the Examiner has mischaracterized the methods used by Hyman *et al.* and Pollack *et al* in their analysis. These papers did not use traditional CGH analysis to identify amplified genes. In Hyman *et al.*, 13,824 cDNA clones were placed on glass slides in a microarray and genomic DNA from breast cancer cell lines and normal human WBCs were hybridized to the cDNA sequences. For expression analysis, RNA from tumor cell lines was hybridized on the same microarrays. The 13,824 arrayed cDNA clones were analyzed for gene expression and gene copy number in 14 breast cancer cell lines. Further, Hyman *et al.* state that "[t]he cDNA/CGH microarray technique enables the direct correlation of copy number and expression data on a gene-by-gene basis throughout the genome." (See page 6242, column 2). Therefore, the analysis performed by Hyman *et al.* was on a gene-by gene basis, and clearly shows that "it is more likely than not" that a gene which is amplified in tumor cells will have increased gene expression.

The Examiner also appears to misunderstand the data presented by Hyman *et al.* The Examiner has asserted that "of the 12,000 transcripts analyzed, a set of 270 was identified in which overexpression was attributable to gene amplification." The Examiner concludes that "[t]his proportion is 2%; the Examiner maintains that 2% does not provide a reasonable expectation that the slight amplification of PRO351 would be correlated with elevated levels of mRNA, much less protein." (Page 7 of the Office Action mailed August 29, 2005). Appellants respectfully submit that the Examiner appears to have misinterpreted the results of Hyman *et al.* Hyman *et al.* chose to do a genome-wide analysis of a large number of genes, most of which, as shown in Figure 2, were not amplified. Accordingly, the 2% number is meaningless, as the low figure mainly results from the fact that only a small percentage of genes are amplified in the first place. The significant figure is not the percentage of genes in the genome that show amplification, but the percentage of amplified genes that demonstrate increased mRNA and protein expression.

The Examiner has further asserted that the Hyman reference "found 44% of *highly* amplified genes showing overexpression at the mRNA level, and 10.5% of highly overexpressed genes being amplified; thus, even at the level of high amplification and high overexpression, the two do not correlate." (Page 7 of the Office Action mailed August 29, 2005). Appellants submit that the 10.5% figure is not relevant to the issue at hand. One of skill in the art would understand that there can be more than one cause of overexpression. The issue is not whether overexpression is always, or even typically caused by gene amplification, but rather, whether gene amplification typically leads to overexpression.

The Examiner's assertion is not consistent with the interpretation Hyman *et al.* themselves place on their data, stating that, "The results illustrate **a considerable influence of copy number on gene expression patterns**." (page 6242. col. 1; emphasis added). In the more detailed discussion of their results, Hyman *et al.* teach that "[u]p to 44% of the highly amplified transcripts (CGH ratio, >2.5) were overexpressed (***i.e.*, belonged to the global upper 7% of expression ratios**) compared with only 6% for genes with normal copy number." (See page 6242, col. 1; emphasis added). These details make it clear that Hyman *et al.* set a highly restrictive standard for considering a gene to be overexpressed; yet almost half of all highly amplified transcripts met even this highly restrictive standard. Therefore, the analysis performed by Hyman *et al.* clearly shows that "it is more likely than not" that a gene which is amplified in tumor cells will have increased gene expression.

In Pollack *et al.*, DNA copy number alteration across 6,691 mapped human genes in 44 predominantly advanced primary breast tumors and 10 breast cancer cell lines was profiled. Pollack *et al.* further state, "Parallel microarray measurements of mRNA levels reveal the remarkable degree to which variation in gene copy number contributes to variation in gene expression in tumor cells." (See Abstract). "Genome-wide, of 117 high-level DNA amplifications (fluorescence ratios >4, and representing 91 different genes), 62% (representing 54 different genes; ...) are found associated with at least moderately elevated mRNA levels (mean-centered fluorescence ratios >2), and 42% (representing 36 different genes) are found associated with comparably highly elevated mRNA levels (mean-centered fluorescence ratios >4)." (See page 12966, column 1). Therefore, the analysis performed by Pollack *et al.* was also

on a gene-by gene basis, and clearly shows that "it is more likely than not" that a gene which is amplified in tumor cells will have increased gene expression.

The Examiner has asserted that Pollack *et al.* is allegedly "limited to highly amplified genes which were not evaluated by the method of the instant specification." (Page 7 of the Office Action mailed August 29, 2005). Appellants respectfully submit that there is no disclosure in Pollack *et al.* which indicates that it is limited to regions showing high amplification. Pollack *et al.* states that their method had the sensitivity to detect 1.5, 2 or 2.5 fold gains in single copy DNA (page 12964), and reports that "on average, a 2-fold change in DNA copy number is associated with a corresponding 1.5-fold change in mRNA levels" (Abstract). As discussed above, the PRO351 gene showed at least 2-fold amplification in ten different lung tumors; thus this is well within the range shown by Pollack *et al.* to produce corresponding changes in mRNA levels.

The Examiner has further asserted that "none of the three papers reported that the research was relevant to identifying probes that can be used as cancer diagnostics." (Page 5 of the Advisory Action mailed October 14, 2004). Appellants respectfully point out that Hyman *et al.* conducted additional studies of one of the genes found to be amplified, HOXB7, and found "**a clinical association between HOXB7 amplification and poor patient prognosis**." (Page 6244, col.1 to col.2; emphasis added). Thus the results of Hyman *et al.* confirm that genes which are amplified in tumors have prognostic utility. The Board's attention is also respectfully directed to the final paragraph of Pollack *et al.*, wherein the authors conclude that "a substantial portion of the phenotypic uniqueness (and, by extension, the heterogeneity in clinical behavior) among patients' tumors may be traceable to underlying variation in DNA copy number." (Page 12698, col. 2). Accordingly, Pollack *et al.* confirm that genes that are amplified in at least one type of tumor are useful as markers for that type of tumor, and for prognostic uses directed to that type of tumor.

Thus these articles, Orntoft, Hyman and Pollack, collectively teach that in general, gene amplification increases mRNA expression.

With respect to the correlation between mRNA expression and protein levels, the Examiner has asserted that the Polakis Declaration is insufficient to overcome the rejection of

claims 58-62 since it is limited to a discussion of data regarding the correlation of mRNA levels and polypeptide levels and not gene amplification levels. The Examiner has asserted that the Declaration does not provide data such that the Examiner can independently draw conclusions. (Page 6 of the Advisory Action mailed October 14, 2004).

Appellants submit that Dr. Polakis' Declaration was presented to support the position that there is a correlation between mRNA levels and polypeptide levels, the correlation between gene amplification and mRNA levels having already been established by the data shown in the Orntoft *et al.*, Hyman *et al.*, and Pollack *et al.* articles. Appellants emphasize that the opinions expressed in the Polakis Declaration, including the quoted statement, are all based on factual findings. Subsequently, antibodies binding to about 30 of these tumor antigens were prepared, and mRNA and protein levels were compared. In approximately 80% of the cases, the researchers found that increases in the level of a particular mRNA correlated with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells. Dr. Polakis' statement that "an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell" is based on factual, experimental findings, clearly set forth in the Declaration. Accordingly, the Declaration is not merely conclusive, and the fact-based conclusions of Dr. Polakis would be considered reasonable and accurate by one skilled in the art.

The case law has clearly established that in considering affidavit evidence, the Examiner must consider all of the evidence of record anew.[18] "After evidence or argument is submitted by the Appellant in response, patentability is determined on the totality of the record, by a preponderance of the evidence with due consideration to persuasiveness of argument"[19] Furthermore, the Federal Court of Appeals held in *In re Alton*, "[W]e are aware of no reason why opinion evidence relating to a fact issue should not be considered by an examiner."[20] Appellants

---

[18] *In re Rinehart*, 531 F.2d 1084, 189 U.S.P.Q. 143 (C.C.P.A. 1976); *In re Piasecki*, 745 F.2d. 1015, 226 U.S.P.Q. 881 (Fed. Cir. 1985).

[19] *In re Alton*, 37 U.S.P.Q.2d 1578, 1584 (Fed. Cir. 1996)(quoting *In re Oetiker*, 977 F.2d 1443, 1445, 24 U.S.P.Q.2d 1443, 1444 (Fed. Cir. 1992)).

[20] *Id.* at 1583.

also respectfully draw the Examiner's attention to the Utility Examination Guidelines[21] which state, "Office personnel must accept an opinion from a qualified expert that is based upon relevant facts whose accuracy is not being questioned; it is improper to disregard the opinion solely because of a disagreement over the significance or meaning of the facts offered." The statement in question from an expert in the field (the Polakis Declaration) states that "it is my considered scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell." Therefore, barring evidence to the contrary regarding the above statement in the Polakis Declaration, this rejection is improper under both the case law and the Utility guidelines.

The Examiner asserts that "there is strong opposing evidence showing that gene amplification is not predictive of increased mRNA levels in normal and cancerous tissues and, in turn, that increased mRNA levels are frequently not predictive of increased polypeptide levels." (Page 6 of the Advisory Action mailed December 7, 2005). In support of this assertion, the Examiner refers to the previously cited references by Pennica *et al.*, Konopka *et al.*, Hu *et al*, Haynes *et al.*, Lian *et al.* and Fessler *et al.*, as well as newly cited references by Chen *et al.*, LaBaer, Gygi *et al.*, and Greenbaum *et al.*

Appellants respectfully submit that, as discussed in detail above, the arguments presented by the Examiner in combination with the Pennica, Konopka, Hu, LaBaer, Chen, Haynes, Gygi, Lian, Fessler, and Greenbaum articles do not provide sufficient reasons to doubt the statements by Appellants that PRO351 has utility. As discussed above, the law does not require the existence of a "necessary" correlation between mRNA and protein levels. Nor does the law require that protein levels be "accurately predicted." According to the authors themselves, the data in the above cited references confirm that there is a general trend between protein expression and transcript levels, which meets the "more likely than not standard" and show that a positive correlation exists between mRNA and protein.

---

[21] Part IIB, 66 Fed. Reg. 1098 (2001).

Taken together, although there are some examples in the scientific art that do not fit within the central dogma of molecular biology that there is a correlation between polypeptide and mRNA levels, these instances are exceptions rather than the rule. In the majority of amplified genes, the teachings in the art, as exemplified by Orntoft *et al.*, Hyman *et al.*, Pollack *et al.*, and the Polakis Declaration, overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO351 gene, that the PRO351 polypeptide is concomitantly overexpressed. Thus, Appellants submit that the PRO351 polypeptide and the claimed antibodies that bind it have utility in the diagnosis of cancer.

E. Even if a *prima facie* case of lack of utility has been established, it should be withdrawn on consideration of the totality of evidence

Even if one assumes *arguendo* that it is more likely than not that there is no correlation between gene amplification and increased mRNA/protein expression, which Appellants submit is **not** true, a polypeptide encoded by a gene that is amplified in cancer would **still** have a specific, substantial, and credible utility. In support, Appellants respectfully draw the Board's attention to page 2 of the Declaration of Dr. Avi Ashkenazi (submitted with the Response filed April 29, 2004) which explains that,

> even when amplification of a cancer marker gene does not result in significant over-expression of the corresponding gene product, this very absence of gene product over-expression still provides significant information for cancer diagnosis and treatment. Thus, if over-expression of the gene product does not parallel gene amplification in certain tumor types but does so in others, then parallel monitoring of gene amplification and gene product over-expression enables more accurate tumor classification and hence better determination of suitable therapy. In addition, absence of over-expression is crucial information for the practicing clinician. If a gene is amplified but the corresponding gene product is not over-expressed, the clinician accordingly will decide not to treat a patient with agents that target that gene product.

Appellants thus submit that simultaneous testing of gene amplification and gene product over-expression enables more accurate tumor classification, even if the gene-product, the protein, is not over-expressed. This leads to better determination of a suitable therapy. Further, as explained in Dr. Ashkenazi's Declaration, absence of over-expression of the protein itself is

crucial information for the practicing clinician. If a gene is amplified in a tumor, but the corresponding gene product is not over-expressed, the clinician will decide not to treat a patient with agents that target that gene product. This not only saves money, but also has the benefit that the patient can avoid exposure to the side effects associated with such agents.

This utility is further supported by the teachings of the article by Hanna and Mornin. (Pathology Associates Medical Laboratories, August (1999); submitted in the IDS filed July 28, 2004). The article teaches that the HER-2/neu gene has been shown to be amplified and/or over-expressed in 10%-30% of invasive breast cancers and in 40%-60% of intraductal breast carcinomas. Further, the article teaches that diagnosis of breast cancer includes testing both the amplification of the HER-2/neu gene (by FISH) as well as the over-expression of the HER-2/neu gene product (by IHC). Even when the protein is not over-expressed, the assay relying on both tests leads to a more accurate classification of the cancer and a more effective treatment of it.

The Examiner has asserted that Hanna *et al.* allegedly "show that gene amplification does not reliably correlate with protein over-expression, and thus the level of polypeptide expression must be tested empirically." (Pages 9-10 of the Office Action mailed August 29, 2005). Appellants respectfully point out that the Examiner appears to have misread Hanna *et al.* Hanna *et al.* clearly state that gene amplification (as measured by FISH) and polypeptide expression (as measured by immunohistochemistry, IHC) are well correlated ("in general, FISH and IHC results correlate well" (Hanna *et al.* p. 1, col. 2)). It is only a subset of tumors which show discordant results. Thus Hanna *et al.* support Appellants' position that it is more likely than not that gene amplification correlates with increased polypeptide expression.

Appellants have clearly shown that the gene encoding the PRO351 polypeptide is amplified in at least ten primary lung tumors. Therefore, the PRO351 gene, similar to the HER-2/neu gene disclosed in Hanna *et al.*, is a tumor associated gene. Furthermore, as discussed above, in the majority of amplified genes, the teachings in the art overwhelmingly show that gene amplification influences gene expression at the mRNA and protein levels. Therefore, one of skill in the art would reasonably expect in this instance, based on the amplification data for the PRO351 gene, that the PRO351 polypeptide is concomitantly overexpressed.

However, even if gene amplification does not result in overexpression of the gene product (*i.e.*, the protein) an analysis of the expression of the protein is useful in determining the course of treatment, as supported by the Ashkenazi Declaration. The Examiner has asserted that "there is no indication that the PRO351 protein levels increase or stay the same. Further research would be needed to determine PRO351 protein levels in cancers showing gene amplification of the PRO351 gene." (Page 4 of the Office Action mailed June 16, 2004). The Examiner appears to view the testing described in the Ashkenazi Declaration and the Hanna paper as experiments involving further characterization of the PRO351 polypeptide itself. In fact, such testing is for the purpose of characterizing not the PRO351 polypeptide, but the tumors in which the gene encoding PRO351 is amplified. The PRO351 polypeptide is therefore useful in tumor categorization, the results of which become an important tool in the hands of a physician enabling the selection of a treatment modality that holds the most promise for the successful treatment of a patient.

For the reasons given above, Appellants respectfully submit that the present specification clearly describes, details and provides a patentable utility for the claimed invention. Accordingly, Appellants respectfully request reconsideration and reversal of the rejections of Claims 58-62 under 35 U.S.C. §101.

ISSUE II: Claims 58-62 satisfy the enablement requirement of 35 U.S.C. §112, first paragraph.

Claims 58-62 stand rejected under 35 U.S.C. §112, first paragraph, allegedly "since the claimed invention is not supported by either a credible, specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art clearly would not know how to use the claimed invention." (Pages 2-3 of the Office Action mailed August 29, 2005).

In this regard, Appellants refer to the arguments and information presented above in response to the outstanding rejection under 35 U.S.C. § 101, wherein those arguments are incorporated by reference herein. Appellants respectfully submit that as described above, the PRO351 polypeptide has utility in the diagnosis of cancer and based on such a utility, one of skill in the art would know exactly how to use the claimed antibodies that bind the PRO351 polypeptide for diagnosis of cancer, without undue experimentation.

Accordingly, Appellants respectfully request reconsideration and reversal of the enablement rejection of Claims 58-62 under 35 U.S.C. §112, first paragraph.

CONCLUSION

For the reasons given above, Appellants submit that the specification discloses at least one patentable utility for the antibodies of Claims 58-62, and that one of ordinary skill in the art would understand how to used the claimed antibodies, for example in the diagnosis of lung tumors. Therefore, Claims 58-62 meet the requirements of 35 U.S.C. §101 and 35 U.S.C. §112, first paragraph.

Accordingly, reversal of all the rejections of Claims 58-62 is respectfully requested.

Please charge any additional fees, including fees for additional extension of time, or credit overpayment to Deposit Account No. **08-1641** (referencing Attorney's Docket No. **39780-2630 P1C11**).

Respectfully submitted,

Date: March 24, 2006

By: _____
Barrie D. Greene (Reg. No. 46,740)

HELLER EHRMAN LLP
275 Middlefield Road
Menlo Park, California 94025-3506
Telephone: (650) 324-7000
Facsimile: (650) 324-0638

8. CLAIMS APPENDIX

Claims on Appeal

58. An isolated antibody that specifically binds to the polypeptide of SEQ ID NO:132.

59. The antibody of Claim 58 which is a monoclonal antibody.

60. The antibody of Claim 58 which is a humanized antibody.

61. An antigen binding fragment of the antibody of Claim 58.

62. The antibody of Claim 28 which is labeled.

9. EVIDENCE APPENDIX

1. Declaration of Avi Ashkenazi, Ph.D. under 37 C.F.R. §1.132; with attached Exhibit A (Curriculum Vitae).

2. Declaration of Paul Polakis, Ph.D. under 37 C.F.R. §1.132.

3. Orntoft, T.F., et al., "Genome-wide Study of Gene Copy Numbers, Transcripts, and Protein Levels in Pairs of Non-Invasive and Invasive Human Transitional Cell Carcinomas," *Molecular & Cellular Proteomics* **1**:37-45 (2002).

4. Hyman, E., et al., "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer," *Cancer Research* **62**:6240-6245 (2002).

5. Pollack, J.R., et al., "Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors," *Proc. Natl. Acad. Sci. USA* **99**:12963-12968 (2002).

6. Hanna, J.S., et al., "HER-2/neu Breast Cancer Predictive Testing," Pathology Associates Medical Laboratories (1999).

7. Declaration of Audrey D. Goddard, Ph.D. under 37 C.F.R. §1.132, with attached Exhibits A-G:

A. Curriculum Vitae of Audrey D. Goddard, Ph.D.

B. Higuchi, R. et al., "Simultaneous amplification and detection of specific DNA sequences," *Biotechnology* **10**:413-417 (1992).

C. Livak, K.J., et al., "Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization," *PCR Methods Appl.* **4**:357-362 (1995).

D. Heid, C.A. et al., "Real time quantitative PCR," *Genome Res.* **6**:986-994 (1996).

E. Pennica, D. et al., "WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors," *Proc. Natl. Acad. Sci. USA* **95**:14717-14722 (1998).

F. Pitti, R.M. et al., "Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer," *Nature* **396**:699-703 (1998).

G. Bieche, I. et al., "Novel approach to quantitative polymerase chain reaction using real-time detection: Application to the detection of gene amplification in breast cancer," *Int. J. Cancer* **78**:661-666 (1998).

8. Pennica, D. et al., "WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors," *Proc. Natl. Acad. Sci. USA* **95**:14717-14722 (1998).

9. Konopka, J.B. et al., "Variable expression of the translocated c-abl oncogene in Philadelphia-chromosome-positive B-lymphoid cell lines from chronic myelogenous leukemia patients," *Proc. Natl. Acad. Sci. USA* **83**:4049-4052 (1986).

10. Haynes, P.A. et al., "Proteome analysis: Biological assay or data archive?" *Electrophoresis* **19**:1862-1871 (1998).

11. Hu, Y. et al., "Analysis of genomic and proteomic data using advanced literature mining," *Journal of Proteome Research* **2**:405-412 (2003).

12. Hittelman, W., "Genetic instability in epithelial tissues at risk for cancer," *Ann. NY Acad Sci.* **952**:1-12 (2001).

13. Fessler, M. B. et al., "A genomic and proteomic analysis of activation of the human neutrophil by lipopolysaccharide and its mediation by p38 mitogen-activated protein kinase," *J. Biol. Chem.* **277**:31291-31302 (2002).

14. Lian, Z. et al., "Genomic and proteomic analysis of the myeloid differentiation program," *Blood* **98**:513-524 (2001).

15. Greenbaum, D. et al., "Comparing protein abundance and mRNA expression levels on a genomic scale," Genome Biol. 4:117.1-117.8- (2003).

16. Chen, G. et al., "Discordant protein and mRNA expression in lung adenocarcinomas," *Mol. Cell. Proteomics* **1**:304-313 (2002).

17. LaBaer, J., "Mining the literature and large datasets," Nature Biotechnology 21:976-977 (2003).

18. Gygi, S. P. et al., "Correlation between protein and mRNA abundance in yeast," *Mol. Cell. Biol.* **19**:1720-1730 (1999).

Item 1 was submitted with Appellants' Response filed April 29, 2004, and acknowledged as having been entered into the record by the Examiner in the Office Action mailed June 16, 2004.

Item 2 was submitted with Appellants' Response filed July 28, 2004, and acknowledged as having been considered by the Examiner in the Advisory Action mailed October 14, 2004.

Items 3-6 were made of record by Appellants in their IDS filed July 28, 2004, and initialed as having been considered by the Examiner on October 12, 2004.

Item 7 was submitted with Appellants' Response filed and acknowledged as having been considered by the Examiner in the Office Action mailed

Items 8-10 were first cited by the Examiner in the Office Action mailed February 24, 2004, and were made of record in the Advisory Action mailed December 7, 2005.

Item 11 was first cited by the Examiner in the Advisory Action of October 14, 2004, and was made of record by the Examiner in the Advisory Action mailed December 7, 2005.

Items 12-14 were made of record by the Examiner in the Office Action mailed August 29, 2005, and again in the Advisory Action mailed December 7, 2005.

Items 15-18 were made of record by the Examiner in the Advisory Action mailed December 7, 2005.

10. RELATED PROCEEDINGS APPENDIX

None.

SV 2192676 v1
3/24/06 10:37 AM (39780.2630)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	: Ashkenazi et al.	Group Art Unit 1647
App. No.	: 09/903,925	CERTIFICATE OF EXPRESS MAILING
Filed	: July 11, 2001	I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to Commissioner of Patents, Washington D.C. 20231 on:
For	: SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS ENCODING THE SAME	(Date)
Examiner	: Hamud, Fozia M	

Commissioner of Patents
P.O. Box 1450
Alexandria, VA 22313-1450

DECLARATION OF AVI ASHKENAZI, Ph.D UNDER 37 C.F.R. § 1.132

I, Avi Ashkenazi, Ph.D. declare and say as follows: -

1. I am Director and Staff Scientist at the Molecular Oncology Department of Genentech, Inc., South San Francisco, CA 94080.

2. I joined Genentech in 1988 as a postdoctoral fellow. Since then, I have investigated a variety of cellular signal transduction mechanisms, including apoptosis, and have developed technologies to modulate such mechanisms as a means of therapeutic intervention in cancer and autoimmune disease. I am currently involved in the investigation of a series of secreted proteins over-expressed in tumors, with the aim to identify useful targets for the development of therapeutic antibodies for cancer treatment.

3. My scientific Curriculum Vitae, including my list of publications, is attached to and forms part of this Declaration (Exhibit A).

4. Gene amplification is a process in which chromosomes undergo changes to contain multiple copies of certain genes that normally exist as a single copy, and is an important factor in the pathophysiology of cancer. Amplification of certain genes (e.g., Myc or Her2/Neu)

gives cancer cells a growth or survival advantage relative to normal cells, and might also provide a mechanism of tumor cell resistance to chemotherapy or radiotherapy.

5. If gene amplification results in over-expression of the mRNA and the corresponding gene product, then it identifies that gene product as a promising target for cancer therapy, for example by the therapeutic antibody approach. Even in the absence of over-expression of the gene product, amplification of a cancer marker gene - as detected, for example, by the reverse transcriptase TaqMan® PCR or the fluorescence *in situ* hybridization (FISH) assays -is useful in the diagnosis or classification of cancer, or in predicting or monitoring the efficacy of cancer therapy. An increase in gene copy number can result not only from intrachromosomal changes but also from chromosomal aneuploidy. It is important to understand that detection of gene amplification can be used for cancer diagnosis even if the determination includes measurement of chromosomal aneuploidy. Indeed, as long as a significant difference relative to normal tissue is detected, it is irrelevant if the signal originates from an increase in the number of gene copies per chromosome and/or an abnormal number of chromosomes.

6. I understand that according to the Patent Office, absent data demonstrating that the increased copy number of a gene in certain types of cancer leads to increased expression of its product, gene amplification data are insufficient to provide substantial utility or well established utility for the gene product (the encoded polypeptide), or an antibody specifically binding the encoded polypeptide. However, even when amplification of a cancer marker gene does not result in significant over-expression of the corresponding gene product, this very absence of gene product over-expression still provides significant information for cancer diagnosis and treatment. Thus, if over-expression of the gene product does not parallel gene amplification in certain tumor types but does so in others, then parallel monitoring of gene amplification and gene product over-expression enables more accurate tumor classification and hence better determination of suitable therapy. In addition, absence of over-expression is crucial information for the practicing clinician. If a gene is amplified but the corresponding gene product is not over-expressed, the clinician accordingly will decide not to treat a patient with agents that target that gene product.

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so

made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

By: Avi Ashkenazi Date: 9/15/03

Avi Ashkenazi, Ph.D.

SV 455281 v1
9/12/03 3:06 PM (39780.7000)

CURRICULUM VITAE

Avi Ashkenazi

July 2003

Personal:

Date of birth:	29 November, 1956
Address:	1456 Tarrytown Street, San Mateo, CA 94402
Phone:	(650) 578-9199 (home); (650) 225-1853 (office)
Fax:	(650) 225-6443 (office)
Email:	aa@gene.com

Education:

1983:	B.S. in Biochemistry, with honors, Hebrew University, Israel
1986:	Ph.D. in Biochemistry, Hebrew University, Israel

Employment:

1983-1986:	Teaching assistant, undergraduate level course in Biochemistry
1985-1986:	Teaching assistant, graduate level course on Signal Transduction
1986 - 1988:	Postdoctoral fellow, Hormone Research Dept., UCSF, and Developmental Biology Dept., Genentech, Inc., with J. Ramachandran
1988 - 1989:	Postdoctoral fellow, Molecular Biology Dept., Genentech, Inc., with D. Capon
1989 - 1993:	Scientist, Molecular Biology Dept., Genentech, Inc.
1994 -1996:	Senior Scientist, Molecular Oncology Dept., Genentech, Inc.
1996-1997:	Senior Scientist and Interim director, Molecular Oncology Dept., Genentech, Inc.
1997-1990:	Senior Scientist and preclinical project team leader, Genentech, Inc.
1999 -2002:	Staff Scientist in Molecular Oncology, Genentech, Inc.
2002-present:	Staff Scientist and Director in Molecular Oncology, Genentech, Inc.

Awards:

1988:	First prize, The Boehringer Ingelheim Award

Editorial:

Editorial Board Member: Current Biology

Associate Editor, Clinical Cancer Research.

Associate Editor, Cancer Biology and Therapy.

Refereed papers:

1. Gertler, A., Ashkenazi, A., and Madar, Z. Binding sites for human growth hormone and ovine and bovine prolactins in the mammary gland and liver of the lactating cow. *Mol. Cell. Endocrinol.* **34**, 51-57 (1984).
2. Gertler, A., Shamay, A., Cohen, N., Ashkenazi, A., Friesen, H., Levanon, A., Gorecki, M., Aviv, H., Hadari, D., and Vogel, T. Inhibition of lactogenic activities of ovine prolactin and human growth hormone (hGH) by a novel form of a modified recombinant hGH. *Endocrinology* **118**, 720-726 (1986).
3. Ashkenazi, A., Madar, Z., and Gertler, A. Partial purification and characterization of bovine mammary gland prolactin receptor. *Mol. Cell. Endocrinol.* **50**, 79-87 (1987).
4. Ashkenazi, A., Pines, M., and Gertler, A. Down-regulation of lactogenic hormone receptors in Nb2 lymphoma cells by cholera toxin. *Biochemistry Internatl.* **14**, 1065-1072 (1987).
5. Ashkenazi, A., Cohen, R., and Gertler, A. Characterization of lactogen receptors in lactogenic hormone-dependent and independent Nb2 lymphoma cell lines. *FEBS Lett.* **210**, 51-55 (1987).
6. Ashkenazi, A., Vogel, T., Barash, I., Hadari, D., Levanon, A., Gorecki, M., and Gertler, A. Comparative study on in vitro and in vivo modulation of lactogenic and somatotropic receptors by native human growth hormone and its modified recombinant analog. *Endocrinology* **121**, 414-419 (1987).
7. Peralta, E., Winslow, J., Peterson, G., Smith, D., Ashkenazi, A., Ramachandran, J., Schimerlik, M., and Capon, D. Primary structure and biochemical properties of an M2 muscarinic receptor. *Science* **236**, 600-605 (1987).
8. Peralta, E. Ashkenazi, A., Winslow, J., Smith, D., Ramachandran, J., and Capon, D. J. Distincnt primary structures, ligand-binding properties and tissue-specific expression of four human muscarinic acetylcholine receptors. *EMBO J.* **6**, 3923-3929 (1987).
9. Ashkenazi, A., Winslow, J., Peralta, E., Peterson, G., Schimerlik, M., Capon, D., and Ramachandran, J. An M2 muscarinic receptor subtype coupled to both adenylyl cyclase and phosphoinositide turnover. *Science* **238**, 672-675 (1987).

10. Pines, M., Ashkenazi, A., Cohen-Chapnik, N., Binder, L., and Gertler, A. Inhibition of the proliferation of Nb2 lymphoma cells by femtomolar concentrations of cholera toxin and partial reversal of the effect by 12-o-tetradecanoyl-phorbol-13-acetate. *J. Cell. Biochem.* **37**, 119-129 (1988).

11. Peralta, E. Ashkenazi, A., Winslow, J. Ramachandran, J., and Capon, D. Differential regulation of PI hydrolysis and adenylyl cyclase by muscarinic receptor subtypes. *Nature* **334**, 434-437 (1988).

12. Ashkenazi., A. Peralta, E., Winslow, J., Ramachandran, J., and Capon, D. Functionally distinct G proteins couple different receptors to PI hydrolysis in the same cell. *Cell* **56**, 487-493 (1989).

13. Ashkenazi, A., Ramachandran, J., and Capon, D. Acetylcholine analogue stimulates DNA synthesis in brain-derived cells via specific muscarinic acetylcholine receptor subtypes. *Nature* **340**, 146-150 (1989).

14. Lammare, D., Ashkenazi, A., Fleury, S., Smith, D., Sekaly, R., and Capon, D. The MHC-binding and gp120-binding domains of CD4 are distinct and separable. *Science* **245**, 743-745 (1989).

15. Ashkenazi., A., Presta, L., Marsters, S., Camerato, T., Rosenthal, K., Fendly, B., and Capon, D. Mapping the CD4 binding site for human immunodefficiency virus type 1 by alanine-scanning mutagenesis. *Proc. Natl. Acad. Sci. USA.* **87**, 7150-7154 (1990).

16. Chamow, S., Peers, D., Byrn, R., Mulkerrin, M., Harris, R., Wang, W., Bjorkman, P., Capon, D., and Ashkenazi, A. Enzymatic cleavage of a CD4 immunoadhesin generates crystallizable, biologically active Fd-like fragments. *Biochemistry* **29**, 9885-9891 (1990).

17. Ashkenazi, A., Smith, D., Marsters, S., Riddle, L., Gregory, T., Ho, D., and Capon, D. Resistance of primary isolates of human immunodefficiency virus type 1 to soluble CD4 is independent of CD4-rgp120 binding affinity. *Proc. Natl. Acad. Sci. USA.* **88**, 7056-7060 (1991).

18. Ashkenazi, A., Marsters, S., Capon, D., Chamow, S., Figari., I., Pennica, D., Goeddel., D., Palladino, M., and Smith, D. Protection against endotoxic shock by a tumor necrosis factor receptor immunoadhesin. *Proc. Natl. Acad. Sci. USA.* **88**, 10535-10539 (1991).

19. Moore, J., McKeating, J., Huang, Y., Ashkenazi, A., and Ho, D. Virions of primary HIV-1 isolates resistant to sCD4 neutralization differ in sCD4 affinity and glycoprotein gp120 retention from sCD4-sensitive isolates. *J. Virol.* **66**, 235-243 (1992).

20. Jin, H., Oksenberg, D., Ashkenazi, A., Peroutka, S., Duncan, A., Rozmahel., R., Yang, Y., Mengod, G., Palacios, J., and O'Dowd, B. Characterization of the human 5-hydroxytryptamine$_{1B}$ receptor. *J. Biol. Chem.* **267**, 5735-5738 (1992).

21. Marsters, A., Frutkin, A., Simpson, N., Fendly, B. and Ashkenazi, A. Identification of cysteine-rich domains of the type 1 tumor necrosis receptor involved in ligand binding. *J. Biol. Chem.* **267**, 5747-5750 (1992).

22. Chamow, S., Kogan, T., Peers, D., Hastings, R., Byrn, R., and Ashkenazi, A. Conjugation of sCD4 without loss of biological activity via a novel carbohydrate-directed cross-linking reagent. *J. Biol. Chem.* **267**, 15916-15922 (1992).

23. Oksenberg, D., Marsters, A., O'Dowd, B., Jin, H., Havlik, S., Peroutka, S., and Ashkenazi, A. A single amino-acid difference confers major pharmacologic variation between human and rodent 5-HT$_{1B}$ receptors. *Nature* **360**, 161-163 (1992).

24. Haak-Frendscho, M., Marsters, S., Chamow, S., Peers, D., Simpson, N., and Ashkenazi, A. Inhibition of interferon γ by an interferon γ receptor immunoadhesin. *Immunology* **79**, 594-599 (1993).

25. Penica, D., Lam, V., Weber, R., Kohr, W., Basa, L., Spellman, M., Ashkenazi, Shire, S., and Goeddel, D. Biochemical characterization of the extracellular domain of the 75-kd tumor necrosis factor receptor. *Biochemistry* **32**, 3131-3138. (1993).

26. Barfod, L., Zheng, Y., Kuang, W., Hart, M., Evans, T., Cerione, R., and Ashkenazi, A. Cloning and expression of a human CDC42 GTPase Activating Protein reveals a functional SH3-binding domain. *J. Biol. Chem.* **268**, 26059-26062 (1993).

27. Chamow, S., Zhang, D., Tan, X., Mhtre, S., Marsters, S., Peers, D., Byrn, R., Ashkenazi, A., and Yunghans, R. A humanized bispecific immunoadhesin-antibody that retargets CD3+ effectors to kill HIV-1-infected cells. *J. Immunol.* **153**, 4268-4280 (1994).

28. Means, R., Krantz, S., Luna, J., Marsters, S., and Ashkenazi, A. Inhibition of murine erythroid colony formation in vitro by iterferon γ and correction by interferon γ receptor immunoadhesin. *Blood* **83**, 911-915 (1994).

29. Haak-Frendscho, M., Marsters, S., Mordenti, J., Gillet, N., Chen, S., andAshkenazi, A. Inhibition of TNF by a TNF receptor immunoadhesin: comparison with an anti-TNF mAb. *J. Immunol.* **152**, 1347-1353 (1994).

30. Chamow, S., Kogan, T., Venuti, M., Gadek, T., Peers, D., Mordenti, J., Shak, S., and Ashkenazi, A. Modification of CD4 immunoadhesin with monomethoxy-PEG aldehyde via reductive alkilation. *Bioconj. Chem.* **5**, 133-140 (1994).

31. Jin, H., Yang, R., Marsters, S., Bunting, S., Wurm, F., Chamow, S., and Ashkenazi, A. Protection against rat endotoxic shock by p55 tumor necrosis factor (TNF) receptor immunoadhesin: comparison to anti-TNF monoclonal antibody. *J. Infect. Diseases* **170**, 1323-1326 (1994).

32. Beck, J., Marsters, S., Harris, R., Ashkenazi, A., and Chamow, S. Generation of soluble interleukin-1 receptor from an immunoadhesin by specific cleavage. *Mol. Immunol.* **31**, 1335-1344 (1994).

33. Pitti, B., Marsters, M., Haak-Frendscho, M., Osaka, G., Mordenti, J., Chamow, S., and Ashkenazi, A. Molecular and biological properties of an interleukin-1 receptor immunoadhesin. *Mol. Immunol.* **31**, 1345-1351 (1994).

34. Oksenberg, D., Havlik, S., Peroutka, S., and Ashkenazi, A. The third intracellular loop of the 5-HT2 receptor specifies effector coupling. *J. Neurochem.* **64**, 1440-1447 (1995).

35. Bach, E., Szabo, S., Dighe, A., Ashkenazi, A., Aguet, M., Murphy, K., and Schreiber, R. Ligand-induced autoregulation of IFN-γ receptor β chain expression in T helper cell subsets. *Science* **270**, 1215-1218 (1995).

36. Jin, H., Yang, R., Marsters, S., Ashkenazi, A., Bunting, S., Marra, M., Scott, R., and Baker, J. Protection against endotoxic shock by bactericidal/permeability-increasing protein in rats. *J. Clin. Invest.* **95**, 1947-1952 (1995).

37. Marsters, S., Penica, D., Bach, E., Schreiber, R., and Ashkenazi, A. Interferon γ signals via a high-affinity multisubunit receptor complex that contains two types of polypeptide chain. *Proc. Natl. Acad. Sci. USA.* **92**, 5401-5405 (1995).

38. Van Zee, K., Moldawer, L., Oldenburg, H., Thompson, W., Stackpole, S., Montegut, W., Rogy, M., Meschter, C., Gallati, H., Schiller, C., Richter, W., Loetcher, H., Ashkenazi, A., Chamow, S., Wurm, F., Calvano, S., Lowry, S., and Lesslauer, W. Protection against lethal E. coli bacteremia in baboons by pretreatment with a 55-kDa TNF receptor-Ig fusion protein, Ro45-2081. *J. Immunol.* **156**, 2221-2230 (1996).

39. Pitti, R., Marsters, S., Ruppert, S., Donahue, C., Moore, A., and Ashkenazi, A. Induction of apoptosis by Apo-2 Ligand, a new member of the tumor necrosis factor cytokine family. *J. Biol. Chem.* **271**, 12687-12690 (1996).

40. Marsters, S., Pitti, R., Donahue, C., Rupert, S., Bauer, K., and Ashkenazi, A. Activation of apoptosis by Apo-2 ligand is independent of FADD but blocked by CrmA. *Curr. Biol.* **6**, 1669-1676 (1996).

41. Marsters, S., Skubatch, M., Gray, C., and Ashkenazi, A. Herpesvirus entry mediator, a novel member of the tumor necrosis factor receptor family, activates the NF-κB and AP-1 transcription factors. *J. Biol. Chem.* **272**, 14029-14032 (1997).

42. Sheridan, J., Marsters, S., Pitti, R., Gurney, A., Skubatch, M., Baldwin, D., Ramakrishnan, L., Gray, C., Baker, K., Wood, W.I., Goddard, A., Godowski, P., and Ashkenazi, A. Control of TRAIL-induced apoptosis by a family of signaling and decoy receptors. *Science* **277**, 818-821 (1997).

43. Marsters, S., Sheridan, J., Pitti, R., Gurney, A., Skubatch, M., Balswin, D., Huang, A., Yuan, J., Goddard, A., Godowski, P., and Ashkenazi, A. A novel receptor for Apo2L/TRAIL contains a truncated death domain. *Curr. Biol.* **7**, 1003-1006 (1997).

44. Marsters, A., Sheridan, J., Pitti, R., Brush, J., Goddard, A., and Ashkenazi, A. Identification of a ligand for the death-domain-containing receptor Apo3. *Curr. Biol.* **8**, 525-528 (1998).

45. Rieger, J., Naumann, U., Glaser, T., Ashkenazi, A., and Weller, M. Apo2 ligand: a novel weapon against malignant glioma? *FEBS Lett.* **427**, 124-128 (1998).

46. Pender, S., Fell, J., Chamow, S., Ashkenazi, A., and MacDonald, T. A p55 TNF receptor immunoadhesin prevents T cell mediated intestinal injury by inhibiting matrix metalloproteinase production. *J. Immunol.* **160**, 4098-4103 (1998).

47. Pitti, R., Marsters, S., Lawrence, D., Roy, Kischkel, F., M., Dowd, P., Huang, A., Donahue, C., Sherwood, S., Baldwin, D., Godowski, P., Wood, W., Gurney, A., Hillan, K., Cohen, R., Goddard, A., Botstein, D., and Ashkenazi, A. Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396**, 699-703 (1998).

48. Mori, S., Marakami-Mori, K., Nakamura, S., Ashkenazi, A., and Bonavida, B. Sensitization of AIDS Kaposi's sarcoma cells to Apo-2 ligand-induced apoptosis by actinomycin D. *J. Immunol.* **162**, 5616-5623 (1999).

49. Gurney, A. Marsters, S., Huang, A., Pitti, R., Mark, M., Baldwin, D., Gray, A., Dowd, P., Brush, J., Heldens, S., Schow, P., Goddard, A., Wood, W., Baker, K., Godowski, P., and Ashkenazi, A. Identification of a new member of the tumor necrosis factor family and its receptor, a human ortholog of mouse GITR. *Curr. Biol.* **9**, 215-218 (1999).

50. Ashkenazi, A., Pai, R., Fong, s., Leung, S., Lawrence, D., Marsters, S., Blackie, C., Chang, L., McMurtrey, A., Hebert, A., DeForge, L., Khoumenis, I., Lewis, D., Harris, L., Bussiere, J., Koeppen, H., Shahrokh, Z., and Schwall, R. Safety and anti-tumor activity of recombinant soluble Apo2 ligand. *J. Clin. Invest.* **104**, 155-162 (1999).

51. Chuntharapai, A., Gibbs, V., Lu, J., Ow, A., Marsters, S., Ashkenazi, A., De Vos, A., Kim, K.J. Determination of residues involved in ligand binding and signal transmissiion in the human IFN-α receptor 2. *J. Immunol.* **163**, 766-773 (1999).

52. Johnsen, A.-C., Haux, J., Steinkjer, B., Nonstad, U., Egeberg, K., Sundan, A., Ashkenazi, A., and Espevik, T. Regulation of Apo2L/TRAIL expression in NK cells – involvement in NK cell-mediated cytotoxicity. *Cytokine* **11**, 664-672 (1999).

53. Roth, W., Isenmann, S., Naumann, U., Kugler, S., Bahr, M., Dichgans, J., Ashkenazi, A., and Weller, M. Eradication of intracranial human malignant glioma xenografts by Apo2L/TRAIL. *Biochem. Biophys. Res. Commun.* **265**, 479-483 (1999).

54. Hymowitz, S.G., Christinger, H.W., Fuh, G., Ultsch, M., O'Connell, M., Kelley, R.F., Ashkenazi, A. and de Vos, A.M. Triggering Cell Death: The Crystal Structure of Apo2L/TRAIL in a Complex with Death Receptor 5. *Molec. Cell* **4**, 563–571 (1999).

55. Hymowitz, S.G., O'Connel, M.P., Utsch, M.H., Hurst, A., Totpal, K., Ashkenazi, A., de Vos, A.M., Kelley, R.F. A unique zinc-binding site revealed by a high-resolution X-ray structure of homotrimeric Apo2L/TRAIL. *Biochemistry* **39**, 633-640 (2000).

56. Zhou, Q., Fukushima, P., DeGraff, W., Mitchell, J.B., Stetler-Stevenson, M., Ashkenazi, A., and Steeg, P.S. Radiation and the Apo2L/TRAIL apoptotic pathway preferentially inhibit the colonization of premalignant human breast cancer cells overexpressing cyclin D1. *Cancer Res.* **60**, 2611-2615 (2000).

57. Kischkel, F.C., Lawrence, D. A., Chuntharapai, A., Schow, P., Kim, J., and Ashkenazi, A. Apo2L/TRAIL-dependent recruitment of endogenous FADD and Caspase-8 to death receptors 4 and 5. *Immunity* **12**, 611-620 (2000).

58. Yan, M., Marsters, S.A., Grewal, I.S., Wang, H., *Ashkenazi, A., and *Dixit, V.M. Identification of a receptor for BlyS demonstrates a crucial role in humoral immunity. *Nature Immunol.* **1**, 37-41 (2000).

59. Marsters, S.A., Yan, M., Pitti, R.M., Haas, P.E., Dixit, V.M., and Ashkenazi, A. Interaction of the TNF homologues BLyS and APRIL with the TNF receptor homologues BCMA and TACI. *Curr. Biol.* **10**, 785-788 (2000).

60. Kischkel, F.C., and Ashkenazi, A. Combining enhanced metabolic labeling with immunoblotting to detect interactions of endogenous cellular proteins. *Biotechniques* **29**, 506-512 (2000).

61. Lawrence, D., Shahrokh, Z., Marsters, S., Achilles, K., Shih, D. Mounho, B., Hillan, K., Totpal, K. DeForge, L., Schow, P., Hooley, J., Sherwood, S., Pai, R., Leung, S., Khan, L., Gliniak, B., Bussiere, J., Smith, C., Strom, S., Kelley, S., Fox, J., Thomas, D., and Ashkenazi, A. Differential hepatocyte toxicity of recombinant Apo2L/TRAIL versions. *Nature Med.* 7, 383-385 (2001).

62. Chuntharapai, A., Dodge, K., Grimmer, K., Schroeder, K., Martsters, S.A., Koeppen, H., Ashkenazi, A., and Kim, K.J. Isotype-dependent inhibition of tumor growth in vivo by monoclonal antibodies to death receptor 4. *J. Immunol.* **166**, 4891-4898 (2001).

63. Pollack, I.F., Erff, M., and Ashkenazi, A. Direct stimulation of apoptotic signaling by soluble Apo2L/tumor necrosis factor-related apoptosis-inducing ligand leads to selective killing of glioma cells. *Clin. Cancer Res.* 7, 1362-1369 (2001).

64. Wang, H., Marsters, S.A., Baker, T., Chan, B., Lee, W.P., Fu, L., Tumas, D., Yan, M., Dixit, V.M., *Ashkenazi, A., and *Grewal, I.S. TACI-ligand interactions are required for T cell activation and collagen-induced arthritis in mice. *Nature Immunol.* 2, 632-637 (2001).

65. Kischkel, F.C., Lawrence, D. A., Tinel, A., Virmani, A., Schow, P., Gazdar, A., Blenis, J., Arnott, D., and Ashkenazi, A. Death receptor recruitment of endogenous caspase-10 and apoptosis initiation in the absence of caspase-8. *J. Biol. Chem.* 276, 46639-46646 (2001).

66. LeBlanc, H., Lawrence, D.A., Varfolomeev, E., Totpal, K., Morlan, J., Schow, P., Fong, S., Schwall, R., Sinicropi, D., and Ashkenazi, A Tumor cell resistance to death receptor induced apoptosis through mutational inactivation of the proapoptotitc Bcl-2 homolog Bax. *Nature Med.* 8, 274-281 (2002).

67. Miller, K., Meng, G., Liu, J., Hurst, A., Hsei, V., Wong, W-L., Ekert, R., Lawrence, D., Sherwood, S., DeForge, L., Gaudreault., Keller, G., Sliwkowski, M., Ashkenazi, A., and Presta, L. Design, Construction, and analyses of multivalent antibodies. *J. Immunol.* 170, 4854-4861 (2003).

68. Varfolomeev, E., Kischkel, F., Martin, F., Wanh, H., Lawrence, D., Olsson, C., Tom, L., Erickson, S., French, D., Schow, P., Grewal, I. and Ashkenazi, A. Immune system development in APRIL knockout mice. Submitted.

Review articles:

1. Ashkenazi, A., Peralta, E., Winslow, J., Ramachandran, J., and Capon, D., J. Functional role of muscarinic acetylcholine receptor subtype diversity. *Cold Spring Harbor Symposium on Quantitative Biology*. **LIII**, 263-272 (1988).
2. Ashkenazi, A., Peralta, E., Winslow, J., Ramachandran, J., and Capon, D. Functional diversity of muscarinic receptor subtypes in cellular signal transduction and growth. *Trends Pharmacol. Sci.* Dec Supplement, 12-21 (1989).
3. Chamow, S., Duliege, A., Ammann, A., Kahn, J., Allen, D., Eichberg, J., Byrn, R., Capon, D., Ward, R., and Ashkenazi, A. CD4 immunoadhesins in anti-HIV therapy: new developments. *Int. J. Cancer* Supplement 7, 69-72 (1992).
4. Ashkenazi, A., Capon, and D. Ward, R. Immunoadhesins. *Int. Rev. Immunol.* **10**, 217-225 (1993).
5. Ashkenazi, A., and Peralta, E. Muscarinic Receptors. In *Handbook of Receptors and Channels*. (S. Peroutka, ed.), CRC Press, Boca Raton, Vol. **I**, p. 1-27, (1994).
6. Krantz, S. B., Means, R. T., Jr., Lina, J., Marsters, S. A., and Ashkenazi, A. Inhibition of erythroid colony formation in vitro by gamma interferon. In *Molecular Biology of Hematopoiesis* (N. Abraham, R. Shadduck, A. Levine F. Takaku, eds.) Intercept Ltd. Paris, Vol. **3**, p. 135-147 (1994).
7. Ashkenazi, A. Cytokine neutralization as a potential therapeutic approach for SIRS and shock. *J. Biotechnology in Healthcare* **1**, 197-206 (1994).
8. Ashkenazi, A., and Chamow, S. M. Immunoadhesins: an alternative to human monoclonal antibodies. *Immunomethods: A companion to Methods in Enzimology* **8**, 104-115 (1995).
9. Chamow, S., and Ashkenazi, A. Immunoadhesins: Principles and Applications. *Trends Biotech.* **14**, 52-60 (1996).
10. Ashkenazi, A., and Chamow, S. M. Immunoadhesins as research tools and therapeutic agents. *Curr. Opin. Immunol.* **9**, 195-200 (1997).
11. Ashkenazi, A., and Dixit, V. Death receptors: signaling and modulation. *Science* **281**, 1305-1308 (1998).
12. Ashkenazi, A., and Dixit, V. Apoptosis control by death and decoy receptors. *Curr. Opin. Cell. Biol.* **11**, 255-260 (1999).

13. Ashkenazi, A. Chapters on Apo2L/TRAIL; DR4, DR5, DcR1, DcR2; and DcR3. Online Cytokine Handbook (www.apnet.com/cytokinereference/).
14. Ashkenazi, A. Targeting death and decoy receptors of the tumor necrosis factor superfamily. *Nature Rev. Cancer* 2, 420-430 (2002).
15. LeBlanc, H. and Ashkenazi, A. Apoptosis signaling by Apo2L/TRAIL. *Cell Death and Differentiation* 10, 66-75 (2003).
16. Almasan, A. and Ashkenazi, A. Apo2L/TRAIL: apoptosis signaling, biology, and potential for cancer therapy. Cytokine and Growth Factor Reviews 14, 337-348 (2003).

Book:

Antibody Fusion Proteins (Chamow, S., and Ashkenazi, A., eds., John Wiley and Sons Inc.) (1999).

Talks:

1. Resistance of primary HIV isolates to CD4 is independent of CD4-gp120 binding affinity. UCSD Symposium, HIV Disease: Pathogenesis and Therapy. Greenelefe, FL, March 1991.
2. Use of immuno-hybrids to extend the half-life of receptors. IBC conference on Biopharmaceutical Halflife Extension. New Orleans, LA, June 1992.
3. Results with TNF receptor Immunoadhesins for the Treatment of Sepsis. IBC conference on Endotoxemia and Sepsis. Philadelphia, PA, June 1992.
4. Immunoadhesins: an alternative to human antibodies. IBC conference on Antibody Engineering. San Diego, CA, December 1993.
5. Tumor necrosis factor receptor: a potential therapeutic for human septic shock. American Society for Microbiology Meeting, Atlanta, GA, May 1993.
6. Protective efficiacy of TNF receptor immunoadhesin vs anti-TNF monoclonal antibody in a rat model for endotoxic shock. 5th International Congress on TNF. Asilomar, CA, May 1994.
7. Interferon-γ signals via a multisubunit receptor complex that contains two types of polypeptide chain. American Association of Immunologists Conference. San Franciso, CA, July 1995.
8. Immunoadhesins: Principles and Applications. Gordon Research Conference on Drug Delivery in Biology and Medicine. Ventura, CA, February 1996.

9. Apo-2 Ligand, a new member of the TNF family that induces apoptosis in tumor cells. Cambridge Symposium on TNF and Related Cytokines in Treatment of Cancer. Hilton-Head, NC, March 1996.
10. Induction of apoptosis by Apo2 Ligand. American Society for Biochemistry and Molecular Biology, Symposium on Growth Factors and Cytokine Receptors. New Orleans, LA, June, 1996.
11. Apo2 ligand, an extracellular trigger of apoptosis. 2nd Clontech Symposium, Palo Alto, CA, October 1996.
12. Regulation of apoptosis by members of the TNF ligand and receptor families. Stanford University School of Medicine, Palo Alto, CA, December 1996.
13. Apo-3: anovel receptor that regulates cell death and inflammation. 4th International Congress on Immune Consequences of Trauma, Shock, and Sepsis. Munich, Germany, March 1997.
14. New members of the TNF ligand and receptor families that regulate apoptosis, inflammation, and immunity. UCLA School of Medicine, LA, CA, March 1997.
15. Immunoadhesins: an alternative to monoclonal antibodies. 5th World Conference on Bispecific Antibodies. Volendam, Holland, June 1997.
16. Control of Apo2L signaling. Cold Spring Harbor Laboratory Symposium on Programmed Cell Death. Cold Spring Harbor, New York. September, 1997.
17. Chairman and speaker, Apoptosis Signaling session. IBC's 4th Annual Conference on Apoptosis. San Diego, CA., October 1997.
18. Control of Apo2L signaling by death and decoy receptors. American Association for the Advancement of Science. Philladelphia, PA, February 1998.
19. Apo2 ligand and its receptors. American Society of Immunologists. San Francisco, CA, April 1998.
20. Death receptors and ligands. 7th International TNF Congress. Cape Cod, MA, May 1998.
21. Apo2L as a potential therapeutic for cancer. UCLA School of Medicine. LA, CA, June 1998.
22. Apo2L as a potential therapeutic for cancer. Gordon Research Conference on Cancer Chemotherapy. New London, NH, July 1998.
23. Control of apoptosis by Apo2L. Endocrine Society Conference, Stevenson, WA, August 1998.
24. Control of apoptosis by Apo2L. International Cytokine Society Conference, Jerusalem, Israel, October 1998.

25. Apoptosis control by death and decoy receptors. American Association for Cancer Research Conference, Whistler, BC, Canada, March 1999.
26. Apoptosis control by death and decoy receptors. American Society for Biochemistry and Molecular Biology Conference, San Francisco, CA, May 1999.
27. Apoptosis control by death and decoy receptors. Gordon Research Conference on Apoptosis, New London, NH, June 1999.
28. Apoptosis control by death and decoy receptors. Arthritis Foundation Research Conference, Alexandria GA, Aug 1999.
29. Safety and anti-tumor activity of recombinant soluble Apo2L/TRAIL. Cold Spring Harbor Laboratory Symposium on Programmed Cell Death. . Cold Spring Harbor, NY, September 1999.
30. The Apo2L/TRAIL system: therapeutic potential. American Association for Cancer Research, Lake Tahoe, NV, Feb 2000.
31. Apoptosis and cancer therapy. Stanford University School of Medicine, Stanford, CA, Mar 2000.
32. Apoptosis and cancer therapy. University of Pennsylvania School of Medicine, Philladelphia, PA, Apr 2000.
33. Apoptosis signaling by Apo2L/TRAIL. International Congress on TNF. Trondheim, Norway, May 2000.
34. The Apo2L/TRAIL system: therapeutic potential. Cap-CURE summit meeting. Santa Monica, CA, June 2000.
35. The Apo2L/TRAIL system: therapeutic potential. MD Anderson Cancer Center. Houston, TX, June 2000.
36. Apoptosis signaling by Apo2L/TRAIL. The Protein Society, 14th Symposium. San Diego, CA, August 2000.
37. Anti-tumor activity of Apo2L/TRAIL. AAPS annual meeting. Indianapolis, IN Aug 2000.
38. Apoptosis signaling and anti-cancer potential of Apo2L/TRAIL. Cancer Research Institute, UC San Francisco, CA, September 2000.
39. Apoptosis signaling by Apo2L/TRAIL. Kenote address, TNF family Minisymposium, NIH. Bethesda, MD, September 2000.
40. Death receptors: signaling and modulation. Keystone symposium on the Molecular basis of cancer. Taos, NM, Jan 2001.
41. Preclinical studies of Apo2L/TRAIL in cancer. Symposium on Targeted therapies in the treatment of lung cancer. Aspen, CO, Jan 2001.

42. Apoptosis signaling by Apo2L/TRAIL. Wiezmann Institute of Science, Rehovot, Israel, March 2001.
43. Apo2L/TRAIL: Apoptosis signaling and potential for cancer therapy. Weizmann Institute of Science, Rehovot, Israel, March 2001.
44. Targeting death receptors in cancer with Apo2L/TRAIL. Cell Death and Disease conference, North Falmouth, MA, Jun 2001.
45. Targeting death receptors in cancer with Apo2L/TRAIL. Biotechnology Organization conference, San Diego, CA, Jun 2001.
46. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Gordon Research Conference on Apoptosis, Oxford, UK, July 2001.
47. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Cleveland Clinic Foundation, Cleveland, OH, Oct 2001.
48. Apoptosis signaling by death receptors: overview. International Society for Interferon and Cytokine Research conference, Cleveland, OH, Oct 2001.
49. Apoptosis signaling by death receptors. American Society of Nephrology Conference. San Francisco, CA, Oct 2001.
50. Targeting death receptors in cancer. Apoptosis: commercial opportunities. San Diego, CA, Apr 2002.
51. Apo2L/TRAIL signaling and apoptosis resistance mechanisms. Kimmel Cancer Research Center, Johns Hopkins University, Baltimore MD. May 2002.
52. Apoptosis control by Apo2L/TRAIL. (Keynote Address) University of Alabama Cancer Center Retreat, Birmingham, Ab. October 2002.
53. Apoptosis signaling by Apo2L/TRAIL. (Session co-chair) TNF international conference. San Diego, CA. October 2002.
54. Apoptosis signaling by Apo2L/TRAIL. Swiss Institute for Cancer Research (ISREC). Lausanne, Swizerland. Jan 2003.
55. Apoptosis induction with Apo2L/TRAIL. Conference on New Targets and Innovative Strategies in Cancer Treatment. Monte Carlo. February 2003.
56. Apoptosis signaling by Apo2L/TRAIL. Hermelin Brain Tumor Center Symposium on Apoptosis. Detroit, MI. April 2003.
57. Targeting apoptosis through death receptors. Sixth Annual Conference on Targeted Therapies in the Treatment of Breast Cancer. Kona, Hawaii. July 2003.
58. Targeting apoptosis through death receptors. Second International Conference on Targeted Cancer Therapy. Washington, DC. Aug 2003.

Issued Patents:

1. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,329,028 (Jul 12, 1994).
2. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,605,791 (Feb 25, 1997).
3. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 5,889,155 (Jul 27, 1999).
4. Ashkenazi, A., APO-2 Ligand. US patent 6,030,945 (Feb 29, 2000).
5. Ashkenazi, A., Chuntharapai, A., Kim, J., APO-2 ligand antibodies. US patent 6, 046, 048 (Apr 4, 2000).
6. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 6,124,435 (Sep 26, 2000).
7. Ashkenazi, A., Chuntharapai, A., Kim, J., Method for making monoclonal and cross-reactive antibodies. US patent 6,252,050 (Jun 26, 2001).
8. Ashkenazi, A. APO-2 Receptor. US patent 6,342,369 (Jan 29, 2002).
9. Ashkenazi, A. Fong, S., Goddard, A., Gurney, A., Napier, M., Tumas, D., Wood, W. A-33 polypeptides. US patent 6,410,708 (Jun 25, 2002).
10. Ashkenazi, A. APO-3 Receptor. US patent 6,462,176 B1 (Oct 8, 2002).
11. Ashkenazi, A. APO-2LI and APO-3 polypeptide antibodies. US patent 6,469,144 B1 (Oct 22, 2002).
12. Ashkenazi, A., Chamow, S. and Kogan, T. Carbohydrate-directed crosslinking reagents. US patent 6,582,928B1 (Jun 24, 2003).

DECLARATION OF PAUL POLAKIS, Ph.D.

I, Paul Polakis, Ph.D., declare and say as follows:

1. I was awarded a Ph.D. by the Department of Biochemistry of the Michigan State University in 1984. My scientific Curriculum Vitae is attached to and forms part of this Declaration (Exhibit A).

2. I am currently employed by Genentech, Inc. where my job title is Staff Scientist. Since joining Genentech in 1999, one of my primary responsibilities has been leading Genentech's Tumor Antigen Project, which is a large research project with a primary focus on identifying tumor cell markers that find use as targets for both the diagnosis and treatment of cancer in humans.

3. As part of the Tumor Antigen Project, my laboratory has been analyzing differential expression of various genes in tumor cells relative to normal cells. The purpose of this research is to identify proteins that are abundantly expressed on certain tumor cells and that are either (i) not expressed, or (ii) expressed at lower levels, on corresponding normal cells. We call such differentially expressed proteins "tumor antigen proteins". When such a tumor antigen protein is identified, one can produce an antibody that recognizes and binds to that protein. Such an antibody finds use in the diagnosis of human cancer and may ultimately serve as an effective therapeutic in the treatment of human cancer.

4. In the course of the research conducted by Genentech's Tumor Antigen Project, we have employed a variety of scientific techniques for detecting and studying differential gene expression in human tumor cells relative to normal cells, at genomic DNA, mRNA and protein levels. An important example of one such technique is the well known and widely used technique of microarray analysis which has proven to be extremely useful for the identification of mRNA molecules that are differentially expressed in one tissue or cell type relative to another. In the course of our research using microarray analysis, we have identified approximately 200 gene transcripts that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. To date, we have generated antibodies that bind to about 30 of the tumor antigen proteins expressed from these differentially expressed gene transcripts and have used these antibodies to quantitatively determine the level of production of these tumor antigen proteins in both human cancer cells and corresponding normal cells. We have then compared the levels of mRNA and protein in both the tumor and normal cells analyzed.

5. From the mRNA and protein expression analyses described in paragraph 4 above, we have observed that there is a strong correlation between changes in the level of mRNA present in any particular cell type and the level of protein

expressed from that mRNA in that cell type. In approximately 80% of our observations we have found that increases in the level of a particular mRNA correlates with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells.

6. Based upon my own experience accumulated in more than 20 years of research, including the data discussed in paragraphs 4 and 5 above and my knowledge of the relevant scientific literature, it is my considered scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell. In fact, it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded protein. While there have been published reports of genes for which such a correlation does not exist, it is my opinion that such reports are exceptions to the commonly understood general rule that increased mRNA levels are predictive of corresponding increased levels of the encoded protein.

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

Dated: 5/07/04

By: Paul Polakis

Paul Polakis, Ph.D.

SV 2031808 v1

CURRICULUM VITAE

PAUL G. POLAKIS
Staff Scientist
Genentech, Inc
1 DNA Way, MS#40
S. San Francisco, CA 94080

EDUCATION:

Ph.D., Biochemistry, Department of Biochemistry,
Michigan State University (1984)

B.S., Biology. College of Natural Science, Michigan State University (1977)

PROFESSIONAL EXPERIENCE:

2002-present	Staff Scientist, Genentech, Inc S. San Francisco, CA
1999- 2002	Senior Scientist, Genentech, Inc., S. San Francisco, CA
1997 -1999	Research Director Onyx Pharmaceuticals, Richmond, CA
1992- 1996	Senior Scientist, Project Leader, Onyx Pharmaceuticals, Richmond, CA
1991-1992	Senior Scientist, Chiron Corporation, Emeryville, CA.
1989-1991	Scientist, Cetus Corporation, Emeryville CA.
1987-1989	Postdoctoral Research Associate, Genentech, Inc., South SanFrancisco, CA.
1985-1987	Postdoctoral Research Associate, Department of Medicine, Duke University Medical Center, Durham, NC

1984-1985	Assistant Professor, Department of Chemistry, Oberlin College, Oberlin, Ohio
1980-1984	Graduate Research Assistant, Department of Biochemistry, Michigan State University East Lansing, Michigan

PUBLICATIONS:

1. Polakis, P G. and Wilson, J. E. 1982 Purification of a Highly Bindable Rat Brain Hexokinase by High Performance Liquid Chromatography. **Biochem. Biophys. Res. Commun**. 107, 937-943.

2. Polakis, P.G. and Wilson, J. E. 1984 Proteolytic Dissection of Rat Brain Hexokinase: Determination of the Cleavage Pattern during Limited Digestion with Trypsin. **Arch. Biochem. Biophys**. 234, 341-352.

3. Polakis, P. G. and Wilson, J. E. 1985 An Intact Hydrophobic N-Terminal Sequence is Required for the Binding Rat Brain Hexokinase to Mitochondria. **Arch. Biochem. Biophys**. 236, 328-337.

4. Uhing, R.J., **Polakis,P.G**. and Snyderman, R. 1987 Isolaton of GTP-binding Proteins from Myeloid HL60 Cells. **J. Biol. Chem**. 262, 15575-15579.

5. Polakis, P.G., Uhing, R.J. and Snyderman, R. 1988 The Formylpeptide Chemoattractant Receptor Copurifies with a GTP-binding Protein Containing a Distinct 40 kDa Pertussis Toxin Substrate. **J. Biol. Chem**. 263, 4969-4979.

6. Uhing, R. J., Dillon, S., **Polakis, P. G.**, Truett, A. P. and Snyderman, R. 1988 Chemoattractant Receptors and Signal Transduction Processes in Cellular and Molecular Aspects of Inflammation (Poste, G. and Crooke, S. T. eds.) pp 335-379.

7. Polakis, P.G., Evans, T. and Snyderman 1989 Multiple Chromatographic Forms of the Formylpeptide Chemoattractant Receptor and their Relationship to GTP-binding Proteins. **Biochem. Biophys. Res. Commun**. 161, 276-283.

8. Polakis, P. G., Snyderman, R. and Evans, T. 1989 Characterization of G25K, a GTP-binding Protein Containing a Novel Putative Nucleotide Binding Domain. **Biochem. Biophys. Res. Comun.** 160, 25-32.

9. Polakis, P., Weber,R.F., Nevins,B., Didsbury, J. Evans,T. and Snyderman, R. 1989 Identification of the ral and rac1 Gene Products, Low Molecular Mass GTP-binding Proteins from Human Platelets. **J. Biol. Chem**. 264, 16383-16389.

10. Snyderman, R., Perianin, A., Evans, T., **Polakis, P.** and Didsbury, J. 1989 G Proteins and Neutrophil Function. In ADP-Ribosylating Toxins and G Proteins: Insights into Signal Transduction. (J. Moss and M. Vaughn, eds.) Amer. Soc. Microbiol. pp. 295-323.

11. Hart, M.J., **Polakis, P.G.**, Evans, T. and Cerrione, R.A. 1990 The Identification and Charaterization of an Epidermal Growth Factor-Stimulated Phosphorylation of a Specific Low Molecular Mass GTP-binding Protein in a Reconstituted Phospholipid Vesicle System. **J. Biol. Chem.** 265, 5990-6001.

12. Yatani, A., Okabe, K., **Polakis, P.** Halenbeck, R. McCormick, F. and Brown, A. M. 1990 ras p21 and GAP Inhibit Coupling of Muscarinic Receptors to Atrial K^+ Channels. **Cell.** 61, 769-776.

13. Munemitsu, S., Innis, M.A., Clark, R., McCormick, F., Ullrich, A. and **Polakis, P.G.** 1990 Molecular Cloning and Expression of a G25K cDNA, the Human Homolog of the Yeast Cell Cycle Gene CDC42. **Mol. Cell. Biol.** 10, 5977-5982.

14. Polakis, P.G. Rubinfeld, B. Evans, T. and McCormick, F. 1991 Purification of Plasma Membrane-Associated GTPase Activating Protein (GAP) Specific for rap-1/krev-1 from HL60 Cells. **Proc. Natl. Acad. Sci. USA** 88, 239-243.

15. Moran, M. F., **Polakis, P.**, McCormick, F., Pawson, T. and Ellis, C. 1991 Protein Tyrosine Kinases Regulate the Phosphorylation, Protein Interactions, Subcellular Distribution, and Activity of p21ras GTPase Activating Protein. **Mol. Cell. Biol.** 11, 1804-1812

16. Rubinfeld, B., Wong, G., Bekesi, E. Wood, A. McCormick, F. and **Polakis, P. G.** 1991 A Synthetic Peptide Corresponding to a Sequence in the GTPase Activating Protein Inhibits $p21^{ras}$ Stimulation and Promotes Guanine Nucleotide Exchange. **Internatl. J. Peptide and Prot. Res.** 38, 47-53.

17. Rubinfeld, B., Munemitsu, S., Clark, R., Conroy, L., Watt, K., Crosier, W., McCormick, F., and **Polakis, P.** 1991 Molecular Cloning of a GTPase Activating Protein Specific for the Krev-1 Protein $p21^{rap1}$. **Cell** 65, 1033-1042.

18. Zhang, K. Papageorge, A., G., Martin, P., Vass, W. C., Olah, Z., **Polakis, P.**, McCormick, F. and Lowy, D, R. 1991 Heterogenous Amino Acids in RAS and Rap1A Specifying Sensitivity to GAP Proteins. **Science** 254, 1630-1634.

19. Martin, G., Yatani, A., Clark, R., **Polakis, P.**, Brown, A. M. and McCormick, F. 1992 GAP Domains Responsible for $p21^{ras}$-dependent Inhibition of Muscarinic Atrial K^+ Channel Currents. **Science** 255, 192-194.

20. McCormick, F., Martin, G. A., Clark, R., Bollag, G. and **Polakis, P** . 1992 Regulation of p21ras by GTPase Activating Proteins. Cold Spring Harbor **Symposia on Quantitative Biology**. Vol. 56, 237-241.

21. Pronk, G. B., **Polakis, P.**, Wong, G., deVries-Smits, A. M., Bos J. L. and McCormick, F. 1992 $p60^{v\text{-}src}$ Can Associate with and Phosphorylate the $p21^{ras}$ GTPase Activating Protein. **Oncogene** 7,389-394.

22. Polakis P. and McCormick, F. 1992 Interactions Between $p21^{ras}$ Proteins and Their GTPase Activating Proteins. In **Cancer Surveys** (Franks, L. M., ed.) 12, 25-42.

23. Wong, G., Muller, O., Clark, R., Conroy, L., Moran, M., **Polakis, P.** and McCormick, F. 1992 Molecular coloning and nucleic acid binding properties of the GAP-associated tyrosine phosphoprotein p62. **Cell** 69, 551-558.

24. Polakis, P., Rubinfeld, B. and McCormick, F. 1992 Phosphorylation of rap1GAP in vivo and by cAMP-dependent Kinase and the Cell Cycle $p34^{cdc2}$ Kinase in vitro. **J. Biol. Chem**. 267, 10780-10785.

25. McCabe, P.C., Haubrauck, H., **Polakis, P.**, McCormick, F., and Innis, M. A. 1992 Functional Interactions Between $p21^{rap1A}$ and Components of the Budding pathway of *Saccharomyces cerevisiae*. **Mol. Cell. Biol**. 12, 4084-4092.

26. Rubinfeld, B., Crosier, W.J., Albert, I., Conroy,L., Clark, R., McCormick, F. and **Polakis, P**. 1992 Localization of the rap1GAP Catalytic Domain and Sites of Phosphorylation by Mutational Analysis. **Mol. Cell . Biol**. 12, 4634-4642.

27. Ando, S., Kaibuchi, K., Sasaki, K., Hiraoka, T., Nishiyama, T., Mizuno, T., Asada, M., Nunoi, H., Matsuda, I., Matsuura, Y., **Polakis, P.**, McCormick, F. and Takai, Y. 1992 Post-translational processing of rac p21s is important both for their interaction with the GDP/GTP exchange proteins and for their activation of NADPH oxidase. **J. Biol. Chem**. 267, 25709-25713.

28. Janoueix-Lerosey, I., **Polakis, P.**, Tavitian, A. and deGunzberg, J. 1992 Regulation of the GTPase activity of the ras-related rap2 protein. **Biochem. Biophys. Res. Commun**. 189, 455-464.

29. Polakis, P. 1993 GAPs Specific for the rap1/Krev-1 Protein. in GTP-binding Proteins: the *ras*-superfamily. (J.C. LaCale and F. McCormick, eds.) 445-452.

30. Polakis, P. and McCormick, F. 1993 Structural requirements for the interaction of $p21^{ras}$ with GAP, exchange factors, and its bological effector target. **J. Biol Chem**. 268, 9157-9160.

31. Rubinfeld, B., Souza, B. Albert, I., Muller, O., Chamberlain, S., Masiarz, F., Munemitsu, S. and **Polakis, P.** 1993 Association of the APC gene product with beta- catenin. **Science** 262, 1731-1734.

32. Weiss, J., Rubinfeld, B., **Polakis, P.**, McCormick, F. Cavenee, W. A. and Arden, K. 1993 The gene for human rap1-GTPase activating protein (rap1GAP) maps to chromosome 1p35-1p36.1. **Cytogenet. Cell Genet**. 66, 18-21.

33. Sato, K. Y., **Polakis, P.**, Haubruck, H., Fasching, C. L., McCormick, F. and Stanbridge, E. J. 1994 Analysis of the tumor suppressor acitvity of the K-rev gene in human tumor cell lines. **Cancer Res**. 54, 552-559.

34. Janoueix-Lerosey, I., Fontenay, M., Tobelem, G., Tavitian, A., **Polakis, P**. and DeGunzburg, J. 1994 Phosphorylation of rap1GAP during the cell cycle. **Biochem. Biophys. Res. Commun**. 202, 967-975

35. Munemitsu, S., Souza, B., Mueller, O., Albert, I., Rubinfeld, B., and **Polakis, P.** 1994 The APC gene product associates with microtubules in vivo and affects their assembly in vitro. **Cancer Res**. 54, 3676-3681.

36. Rubinfeld, B. and **Polakis, P.** 1995 Purification of baculovirus produced rap1GAP. **Methods Enz.** 255,31

37. Polakis, P. 1995 Mutations in the APC gene and their implications for protein structure and function. **Current Opinions in Genetics and Development** 5, 66-71

38. Rubinfeld, B., Souza, B., Albert, I., Munemitsu, S. and **Polakis P.** 1995 The APC protein and E-cadherin form similar but independent complexes with α-catenin, β-catenin and Plakoglobin. **J. Biol. Chem.** 270, 5549-5555

39. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B., and **Polakis, P.** 1995 Regulation of intracellular β-catenin levels by the APC tumor suppressor gene. **Proc. Natl. Acad. Sci.** 92, 3046-3050.

40. Lock, P., Fumagalli, S., **Polakis, P.** McCormick, F. and Courtneidge, S. A. 1996 The human p62 cDNA encodes Sam68 and not the rasGAP-associated p62 protein. **Cell** 84, 23-24.

41. Papkoff, J., Rubinfeld, B., Schryver, B. and **Polakis, P.** 1996 Wnt-1 regulates free pools of catenins and stabilizes APC-catenin complexes. **Mol. Cell. Biol.** 16, 2128-2134.

42. Rubinfeld, B., Albert, I., Porfiri, E., Fiol, C., Munemitsu, S. and **Polakis, P.** 1996 Binding of GSK3β to the APC-β-catenin complex and regulation of complex assembly. **Science** 272, 1023-1026.

43. Munemitsu, S., Albert, I., Rubinfeld, B. and **Polakis, P.** 1996 Deletion of amino-terminal structure stabilizes β-catenin in vivo and promotes the hyperphosphorylation of the APC tumor suppressor protein. **Mol. Cell. Biol.**16, 4088-4094.

44. Hart, M. J., Callow, M. G., Sousa, B. and **Polakis P.** 1996 IQGAP1, a calmodulin binding protein witha rasGAP related domain, is a potential effector for cdc42Hs. **EMBO J.** 15, 2997-3005.

45. Nathke, I. S., Adams, C. L., **Polakis, P.**, Sellin, J. and Nelson, W. J. 1996 The adenomatous polyposis coli (APC) tumor suppressor protein is localized to plasma membrane sites involved in active epithelial cell migration. **J. Cell. Biol.** 134, 165-180.

46. Hart, M. J., Sharma, S., elMasry, N., Qui, R-G., McCabe, P., **Polakis, P.** and Bollag, G. 1996 Identification of a novel guanine nucleotide exchange factor for the rho GTPase. **J. Biol. Chem.** 271, 25452.

47. Thomas JE, Smith M, Rubinfeld B, Gutowski M, Beckmann RP, and **Polakis P.** 1996 Subcellular localization and analysis of apparent 180-kDa and 220-kDa proteins of the breast cancer susceptibility gene, BRCA1. **J. Biol. Chem.** 1996 271, 28630-28635

48. Hayashi, S., Rubinfeld, B., Souza, B., **Polakis, P.**, Wieschaus, E., and Levine, A. 1997 A Drosophila homolog of the tumor suppressor adenomatous polyposis coli

down-regulates β -catenin but its zygotic expression is not essential for the regulation of armadillo. **Proc. Natl. Acad. Sci.** 94, 242-247.

49. Vleminckx, K., Rubinfeld, B., **Polakis, P.** and Gumbiner, B. 1997 The APC tumor suppressor protein induces a new axis in Xenopus embryos. **J. Cell. Biol.** 136, 411-420.

50. Rubinfeld, B., Robbins, P., El-Gamil, M., Albert, I., Porfiri, P. and **Polakis, P.** 1997 Stabilization of β-catenin by genetic defects in melanoma cell lines. **Science** 275, 1790-1792.

51. **Polakis, P**. The adenomatous polyposis coli (APC) tumor suppressor. 1997 **Biochem. Biophys. Acta**, 1332, F127-F147.

52. Rubinfeld, B., Albert, I., Porfiri, E., Munemitsu, S., and **Polakis, P** 1997 Loss of β-catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene. **Cancer Res**. 57, 4624-4630.

53. Porfiri, E., Rubinfeld, B., Albert, I., Hovanes. K., Waterman, M., and **Polakis, P.** 1997 Induction of a β-catenin-LEF-1 complex by wnt-1 and transforming mutants of β-catenin. **Oncogene** 15, 2833-2839.

54. Thomas JE, Smith M, Tonkinson JL, Rubinfeld B, and **Polakis P.,** 1997 Induction of phosphorylation on BRCA1 during the cell cycle and after DNA damage. **Cell Growth Differ**. 8, 801-809.

55. Hart, M., de los Santos, R., Albert, I., Rubinfeld, B., and **Polakis P.**, 1998 Down regulation of β-catenin by human Axin and its association with the adenomatous polyposis coli (APC) tumor suppressor, β-catenin and glycogen synthase kinase 3β. **Current Biology 8,** 573-581.

56. Polakis, P. 1998 The oncogenic activation of β-catenin. **Current Opinions in Genetics and Development** 9, 15-21

57. Matt Hart, Jean-Paul Concordet, Irina Lassot, Iris Albert, Rico del los Santos, Herve Durand, Christine Perret, Bonnee Rubinfled, Florence Margottin, Richard Benarous and **Paul Polakis**. 1999 The F-box protein β–TrCP associates with phosphorylated β–catenin and regulates its activity in the cell. **Current Biology** 9, 207-10.

58. Howard C. Crawford, Barbara M. Fingleton, Bonnee Rubinfeld, **Paul Polakis** and Lynn M. Matrisian 1999 The metalloproteinase matrilysin is a target of β–catenin transactivation in intestinal tumours. **Oncogene** 18, 2883-91.

59. Meng J, Glick JL, **Polakis P**, Casey PJ. 1999 Functional interaction between Galpha(z) and Rap1GAP suggests a novel form of cellular cross-talk. **J Biol Chem**. 17, 36663-9

60. Vijayasurian Easwaran, Virginia Song, **Paul Polakis** and Steve Byers 1999 The ubiquitin-proteosome pathway and serine kinase activity modulate APC mediated regulation of β-catenin-LEF signaling. **J. Biol. Chem**. 274(23):16641-5.

61 Polakis P, Hart M and Rubinfeld B. 1999 Defects in the regulation of beta-catenin in colorectal cancer. **Adv Exp Med Biol**. 470, 23-32

62 Shen Z, Batzer A, Koehler JA, **Polakis P**, Schlessinger J, Lydon NB, Moran MF. 1999 Evidence for SH3 domain directed binding and phosphorylation of Sam68 by Src. **Oncogene**. 18, 4647-53

64. Thomas GM, Frame S, Goedert M, Nathke I, **Polakis P**, Cohen P. 1999 A GSK3- binding peptide from FRAT1 selectively inhibits the GSK3-catalysed phosphorylation of axin and beta-catenin. **FEBS Lett**. 458, 247-51.

65. Peifer M, **Polakis P**. 2000 Wnt signaling in oncogenesis and embryogenesis--a look outside the nucleus. **Science** 287,1606-9.

66. **Polakis P.** 2000 Wnt signaling and cancer. **Genes Dev**;14, 1837-1851.

67. Spink KE, **Polakis P**, Weis WI 2000 Structural basis of the Axin-adenomatous polyposis coli interaction. **EMBO J** 19, 2270-2279.

68. Szeto , W., Jiang, W., Tice, D.A., Rubinfeld, B., Hollingshead, P.G., Fong, S.E., Dugger, D.L., Pham, T., Yansura, D.E., Wong, T.A., Grimaldi, J.C., Corpuz, R.T., Singh J.S., Frantz, G.D., Devaux, B., Crowley, C.W., Schwall, R.H., Eberhard, D.A., Rastelli, L., **Polakis, P.** and Pennica, D. 2001 Overexpression of the Retinoic Acid-Responsive Gene Stra6 in Human Cancers and its Synergistic Induction by Wnt-1 and Retinoic Acid. **Cancer Res** 61, 4197-4204.

69. Rubinfeld B, Tice DA, **Polakis P**. 2001 Axin dependent phosphorylation of the adenomatous polyposis coli protein mediated by casein kinase 1 epsilon. **J Biol Chem** 276, 39037-39045.

70. **Polakis P**. 2001 More than one way to skin a catenin. **Cell** 2001 105, 563-566.

71. Tice DA, Soloviev I, **Polakis P**. 2002 Activation of the Wnt Pathway Interferes withSerum Response Element-driven Transcription of Immediate Early Genes. **J Biol. Chem**. 277, 6118-6123.

72. Tice DA, Szeto W, Soloviev I, Rubinfeld B, Fong SE, Dugger DL, Winer J,

Williams PM, Wieand D, Smith V, Schwall RH, Pennnica D, **Polakis P.** 2002 Synergistic activation of tumor antigens by wnt-1 signaling and retinoic acid revealed by gene expression profiling. **J Biol Chem**. 277,14329-14335.

73. **Polakis, P.** 2002 Casein kinase I: A wnt'er of disconnect. **Curr. Biol.** 12, R499.

74. Mao,W. , Luis, E., Ross, S., Silva, J., Tan, C., Crowley, C., Chui, C., Franz, G., Senter, P., Koeppen, H., **Polakis, P.** 2004 EphB2 as a therapeutic antibody drug target for the treatment of colorectal cancer. **Cancer Res.** 64, 781-788.

75. Shibamoto, S., Winer, J., Williams, M., Polakis, P. 2003 A Blockade in Wnt signaling is activated following the differentiation of F9 teratocarcinoma cells. **Exp. Cell Res.** 29211-20.

76. Zhang Y, Eberhard DA, Frantz GD, Dowd P, Wu TD, Zhou Y, Watanabe C, Luoh SM, **Polakis P,** Hillan KJ, Wood WI, Zhang Z. 2004 GEPIS--quantitative gene expression profiling in normal and cancer tissues. **Bioinformatics**, April 8

Genome-wide Study of Gene Copy Numbers, Transcripts, and Protein Levels in Pairs of Non-invasive and Invasive Human Transitional Cell Carcinomas*

Torben F. Ørntoft‡§, Thomas Thykjaer¶, Frederic M. Waldman‖, Hans Wolf**, and Julio E. Celis‡‡

Gain and loss of chromosomal material is characteristic of bladder cancer, as well as malignant transformation in general. The consequences of these changes at both the transcription and translation levels is at present unknown partly because of technical limitations. Here we have attempted to address this question in pairs of non-invasive and invasive human bladder tumors using a combination of technology that included comparative genomic hybridization, high density oligonucleotide array-based monitoring of transcript levels (5600 genes), and high resolution two-dimensional gel electrophoresis. The results showed that there is a gene dosage effect that in some cases superimposes on other regulatory mechanisms. This effect depended ($p < 0.015$) on the magnitude of the comparative genomic hybridization change. In general (18 of 23 cases), chromosomal areas with more than 2-fold gain of DNA showed a corresponding increase in mRNA transcripts. Areas with loss of DNA, on the other hand, showed either reduced or unaltered transcript levels. Because most proteins resolved by two-dimensional gels are unknown it was only possible to compare mRNA and protein alterations in relatively few cases of well focused abundant proteins. With few exceptions we found a good correlation ($p < 0.005$) between transcript alterations and protein levels. The implications, as well as limitations, of the approach are discussed. ***Molecular & Cellular Proteomics 1:37–45, 2002.***

Aneuploidy is a common feature of most human cancers (1), but little is known about the genome-wide effect of this phenomenon at both the transcription and translation levels. High throughput array studies of the breast cancer cell line BT474 has suggested that there is a correlation between DNA copy numbers and gene expression in highly amplified areas (2), and studies of individual genes in solid tumors have revealed a good correlation between gene dose and mRNA or protein levels in the case of c-erb-B2, *cyclin d1*, *ems1*, and N-myc (3–5). However, a high cyclin D1 protein expression has been observed without simultaneous amplification (4), and a low level of c-myc copy number increase was observed without concomitant c-myc protein overexpression (6).

In human bladder tumors, karyotyping, fluorescent *In situ* hybridization, and comparative genomic hybridization (CGH)[1] have revealed chromosomal aberrations that seem to be characteristic of certain stages of disease progression. In the case of non-invasive pTa transitional cell carcinomas (TCCs), this includes loss of chromosome 9 or parts of it, as well as loss of Y in males. In minimally invasive pT1 TCCs, the following alterations have been reported: 2q−, 11p−, 1q+, 11q13+, 17q+, and 20q+ (7–12). It has been suggested that these regions harbor tumor suppressor genes and oncogenes; however, the large chromosomal areas involved often contain many genes, making meaningful predictions of the functional consequences of losses and gains very difficult.

In this investigation we have combined genome-wide technology for detecting genomic gains and losses (CGH) with gene expression profiling techniques (microarrays and proteomics) to determine the effect of gene copy number on transcript and protein levels in pairs of non-invasive and invasive human bladder TCCs.

EXPERIMENTAL PROCEDURES

Material—Bladder tumor biopsies were sampled after informed consent was obtained and after removal of tissue for routine pathology examination. By light microscopy tumors 335 and 532 were staged by an experienced pathologist as pTa (superficial papillary),

From the ‡Department of Clinical Biochemistry, Molecular Diagnostic Laboratory and **Department of Urology, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark, ¶AROS Applied Biotechnology ApS, Gustav Wiedsvej 10, DK-8000 Aarhus C, Denmark, ‖UCSF Cancer Center and Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0808, and ‡‡Institute of Medical Biochemistry and Danish Centre for Human Genome Research, Ole Worms Allé 170, Aarhus University, DK-8000 Aarhus C, Denmark

Received, September 26, 2001, and in revised form, November 7, 2001

Published, MCP Papers in Press, November 13, 2001, DOI 10.1074/mcp.M100019-MCP200

[1] The abbreviations used are: CGH, comparative genomic hybridization; TCC, transitional cell carcinoma; LOH, loss of heterozygosity; PA-FABP, psoriasis-associated fatty acid-binding protein; 2D, two-dimensional.

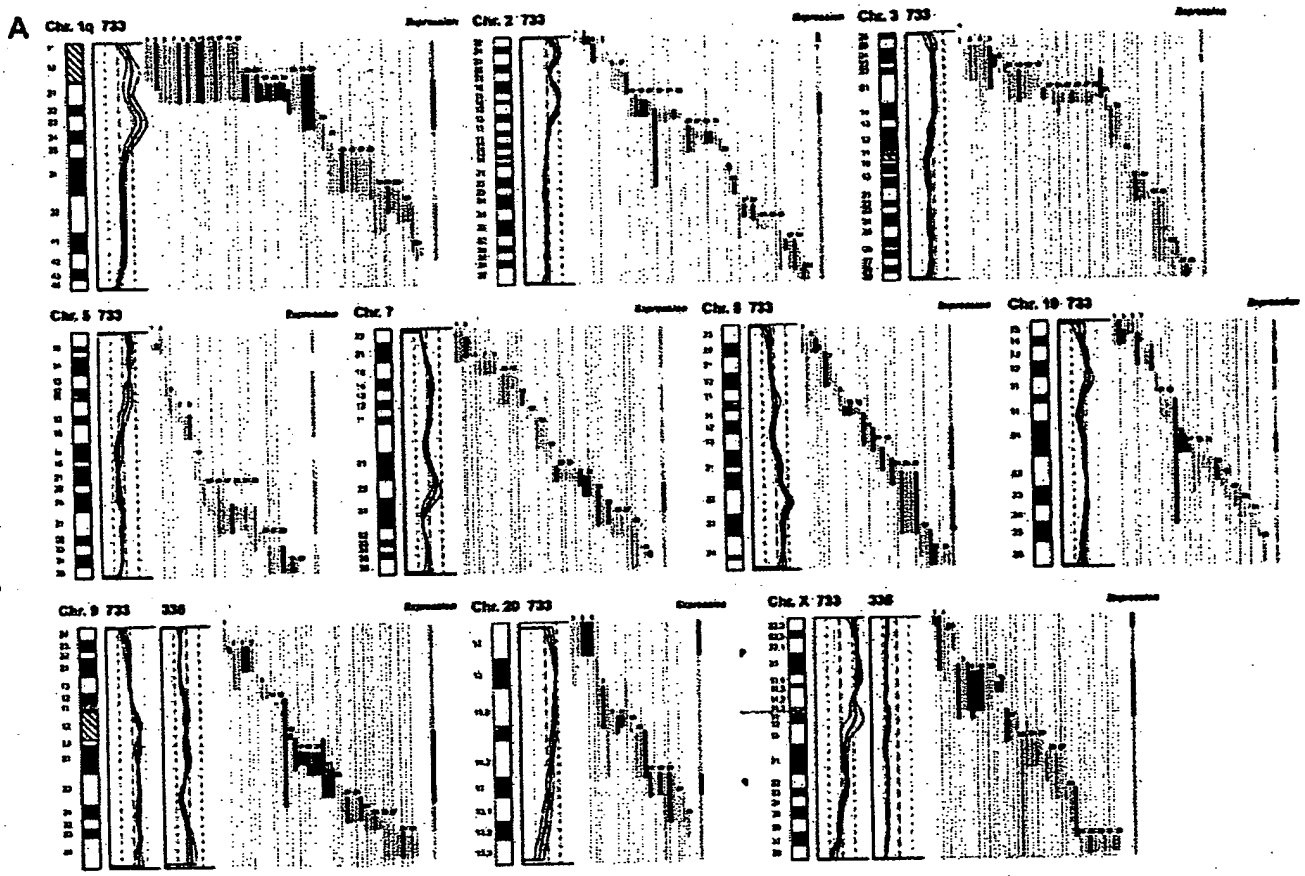


FIG. 1. **DNA copy number and mRNA expression level.** Shown from *left* to *right* are chromosome (*Chr.*), CGH profiles, gene location and expression level of specific genes, and overall expression level along the chromosome. *A*, expression of mRNA in invasive tumor 733 as compared with the non-invasive counterpart tumor 335. *B*, expression of mRNA in invasive tumor 827 compared with the non-invasive counterpart tumor 532. The average fluorescent signal ratio between tumor DNA and normal DNA is shown along the length of the chromosome (*left*). The *bold curve* in the ratio profile represents a mean of four chromosomes and is surrounded by *thin curves* indicating one standard deviation. The *central vertical line* (*broken*) indicates a ratio value of 1 (no change), and the *vertical lines* next to it (*dotted*) indicate a ratio of 0.5 (*left*) and 2.0 (*right*). In chromosomes where the non-invasive tumor 335 used for comparison showed alterations in DNA content, the ratio profile of that chromosome is shown to the *right* of the invasive tumor profile. The *colored bars* represents one gene each, identified by the running *numbers above* the *bars* (the name of the gene can be seen at www.MDL.DK/sdata.html). The *bars* indicate the purported location of the gene, and the *colors* indicate the expression level of the gene in the invasive tumor compared with the non-invasive counterpart; >2-fold increase (*black*), >2-fold decrease (*blue*), no significant change (*orange*). The *bar* to the *far right*, entitled *Expression* shows the resulting change in expression along the chromosome; the *colors* indicate that at least half of the genes were up-regulated (*black*), at least half of the genes down-regulated (*blue*), or more than half of the genes are unchanged (*orange*). If a gene was absent in one of the samples and present in another, it was regarded as more than a 2-fold change. A 2-fold level was chosen as this corresponded to one standard deviation in a double determination of ~1800 genes. Centromeres and heterochromatic regions were excluded from data analysis.

grade I and II, respectively, tumors 733 and 827 were staged as pT1 (invasive into submucosa), 733 was staged as solid, and 827 was staged as papillary, both grade III.

mRNA Preparation—Tissue biopsies, obtained fresh from surgery, were embedded immediately in a sodium-guanidinium thiocyanate solution and stored at −80 °C. Total RNA was isolated using the RNAzol B RNA isolation method (WAK-Chemie Medical GMBH). poly(A)$^+$ RNA was isolated by an oligo(dT) selection step (Oligotex mRNA kit; Qiagen).

cRNA Preparation—1 μg of mRNA was used as starting material. The first and second strand cDNA synthesis was performed using the SuperScript® choice system (Invitrogen) according to the manufacturer's instructions but using an oligo(dT) primer containing a T7 RNA polymerase binding site. Labeled cRNA was prepared using the MEGAscrip® *in vitro* transcription kit (Ambion). Biotin-labeled CTP and UTP (Enzo) was used, together with unlabeled NTPs in the reaction. Following the *in vitro* transcription reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen).

Array Hybridization and Scanning—Array hybridization and scanning was modified from a previous method (13). 10 μg of cRNA was fragmented at 94 °C for 35 min in buffer containing 40 mM Tris acetate, pH 8.1, 100 mM KOAc, 30 mM MgOAc. Prior to hybridization, the fragmented cRNA in a 6× SSPE-T hybridization buffer (1 M NaCl, 10 mM Tris, pH 7.6, 0.005% Triton), was heated to 95 °C for 5 min, subsequently cooled to 40 °C, and loaded onto the Affymetrix probe array cartridge. The probe array was then incubated for 16 h at 40 °C at constant rotation (60 rpm). The probe array was exposed to 10 washes in 6× SSPE-T at 25 °C followed by 4 washes in 0.5× SSPE-T at 50 °C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, 10 μg/ml (Molecular Probes) in 6× SSPE-T

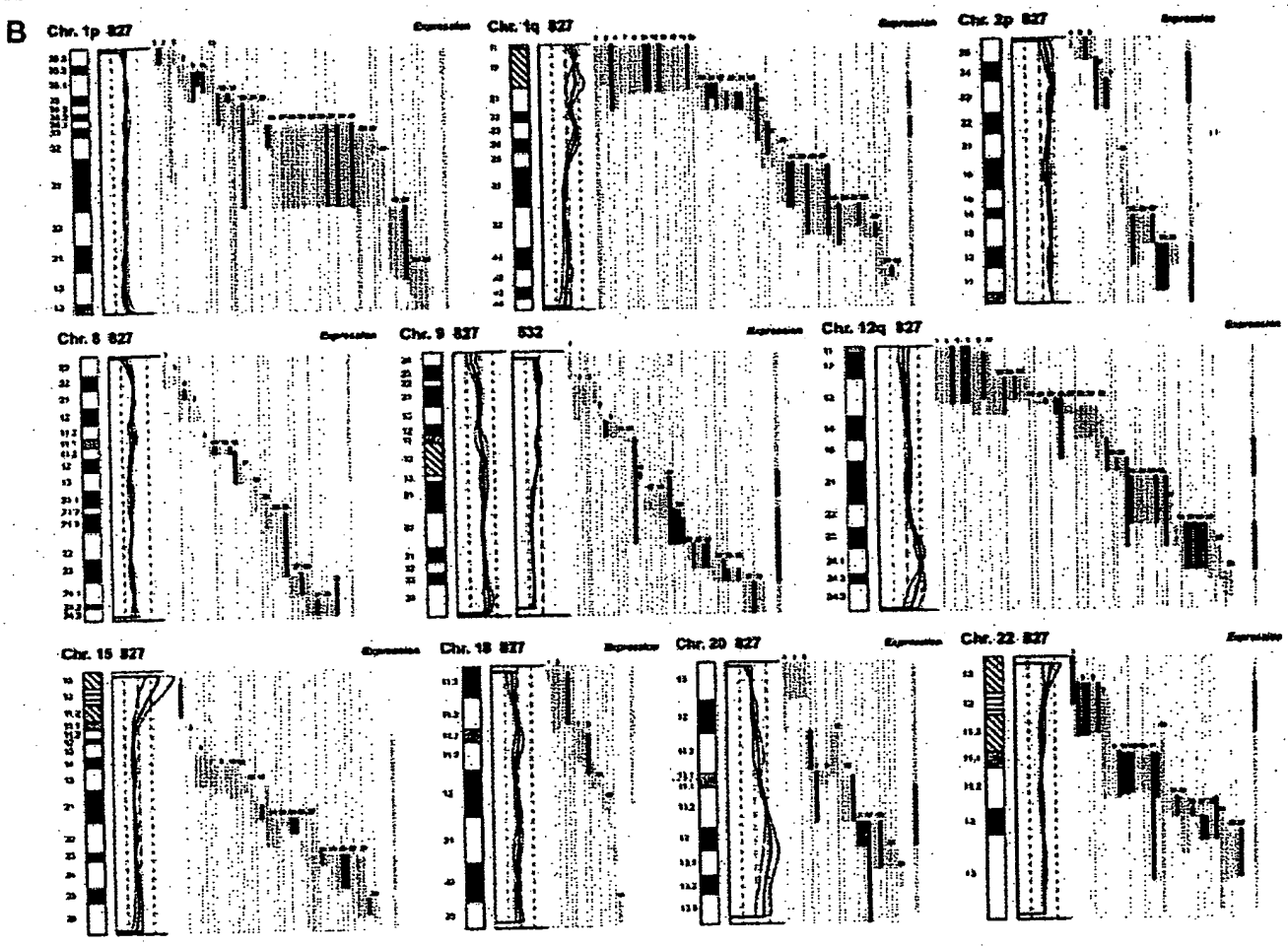


FIG. 1—*continued*

for 30 min at 25 °C followed by 10 washes in 6× SSPE-T at 25 °C. The probe arrays were scanned at 560 nm using a confocal laser scanning microscope (made for Affymetrix by Hewlett-Packard). The readings from the quantitative scanning were analyzed by Affymetrix gene expression analysis software.

Microsatellite Analysis—Microsatellite Analysis was performed as described previously (14). Microsatellites were selected by use of www.ncbi.nlm.nih.gov/genemap98, and primer sequences were obtained from the genome data base at www.gdb.org. DNA was extracted from tumor and blood and amplified by PCR in a volume of 20 μl for 35 cycles. The amplicons were denatured and electrophoresed for 3 h in an ABI Prism 377. Data were collected in the Gene Scan program for fragment analysis. Loss of heterozygosity was defined as less than 33% of one allele detected in tumor amplicons compared with blood.

Proteomic Analysis—TCCs were minced into small pieces and homogenized in a small glass homogenizer in 0.5 ml of lysis solution. Samples were stored at −20 °C until use. The procedure for 2D gel electrophoresis has been described in detail elsewhere (15, 16). Gels were stained with silver nitrate and/or Coomassie Brilliant Blue. Proteins were identified by a combination of procedures that included microsequencing, mass spectrometry, two-dimensional gel Western immunoblotting, and comparison with the master two-dimensional gel image of human keratinocyte proteins; see biobase.dk/cgi-bin/celis.

CGH—Hybridization of differentially labeled tumor and normal DNA to normal metaphase chromosomes was performed as described previously (10). Fluorescein-labeled tumor DNA (200 ng), Texas Red-labeled reference DNA (200 ng), and human Cot-1 DNA (20 μg) were denatured at 37 °C for 5 min and applied to denatured normal metaphase slides. Hybridization was at 37 °C for 2 days. After washing, the slides were counterstained with 0.15 μg/ml 4,6-diamidino-2-phenylindole in an anti-fade solution. A second hybridization was performed for all tumor samples using fluorescein-labeled reference DNA and Texas Red-labeled tumor DNA (inverse labeling) to confirm the aberrations detected during the initial hybridization. Each CGH experiment also included a normal control hybridization using fluorescein- and Texas Red-labeled normal DNA. Digital image analysis was used to identify chromosomal regions with abnormal fluorescence ratios, indicating regions of DNA gains and losses. The average green:red fluorescence intensity ratio profiles were calculated using four images of each chromosome (eight chromosomes total) with normalization of the green:red fluorescence intensity ratio for the entire metaphase and background correction. Chromosome identification was performed based on 4,6-diamidino-2-phenylindole banding patterns. Only images showing uniform high intensity fluorescence with minimal background staining were analyzed. All centromeres, p arms of acrocentric chromosomes, and heterochromatic regions were excluded from the analysis.

RESULTS

Comparative Genomic Hybridization—The CGH analysis identified a number of chromosomal gains and losses in the

TABLE I

Correlation between alterations detected by CGH and by expression monitoring

Top, CGH used as independent variable (if CGH alteration – what expression ratio was found); bottom, altered expression used as independent variable (if expression alteration – what CGH deviation was found).

CGH alterations	Tumor 733 vs. 335 Expression change clusters	Concordance	CGH alterations	Tumor 827 vs. 532 Expression change clusters	Concordance
13 Gain	10 Up-regulation 0 Down-regulation 3 No change	77%	10 Gain	8 Up-regulation 0 Down-regulation 2 No change	80%
10 Loss	1 Up-regulation 5 Down-regulation 4 No change	50%	12 Loss	3 Up-regulation 2 Down regulation 7 No change	17%

Expression change clusters	Tumor 733 vs. 335 CGH alterations	Concordance	Expression change clusters	Tumor 827 vs. 532 CGH alterations	Concordance
16 Up-regulation	11 Gain 2 Loss 3 No change	69%	17 Up-regulation	10 Gain 5 Loss 2 No change	59%
21 Down-regulation	1 Gain 8 Loss 12 No change	38%	9 Down-regulation	0 Gain 3 Loss 6 No change	33%
15 No change	3 Gain 3 Loss 9 No change	60%	21 No change	1 Gain 3 Loss 17 No change	81%

two invasive tumors (stage pT1, TCCs 733 and 827), whereas the two non-invasive papillomas (stage pTa, TCCs 335 and 532) showed only 9p–, 9q22-q33–, and X–, and 7+, 9q–, and Y–, respectively. Both invasive tumors showed changes (1q22–24+, 2q14.1-qter–, 3q12-q13.3–, 6q12-q22–, 9q34+, 11q12-q13+, 17+, and 20q11.2-q12+) that are typical for their disease stage, as well as additional alterations, some of which are shown in Fig. 1. Areas with gains and losses deviated from the normal copy number to some extent, and the average numerical deviation from normal was 0.4-fold in the case of TCC 733 and 0.3-fold for TCC 827. The largest changes, amounting to at least a doubling of chromosomal content, were observed at 1q23 in TCC 733 (Fig. 1*A*) and 20q12 in TCC 827 (Fig. 1*B*).

mRNA Expression in Relation to DNA Copy Number—The mRNA levels from the two invasive tumors (TCCs 827 and 733) were compared with the two non-invasive counterparts (TCCs 532 and 335). This was done in two separate experiments in which we compared TCCs 733 to 335 and 827 to 532, respectively, using two different scaling settings for the arrays to rule out scaling as a confounding parameter. Approximately 1,800 genes that yielded a signal on the arrays were searched in the Unigene and Genemap data bases for chromosomal location, and those with a known location (1096) were plotted as bars covering their purported locus. In that way it was possible to construct a graphic presentation of DNA copy number and relative mRNA levels along the individual chromosomes (Fig. 1).

For each mRNA a ratio was calculated between the level in the invasive *versus* the non-invasive counterpart. Bars, which represent chromosomal location of a gene, were color-coded according to the expression ratio, and only differences larger than 2-fold were regarded as informative (Fig. 1). The density of genes along the chromosomes varied, and areas containing only one gene were excluded from the calculations. The resolution of the CGH method is very low, and some of the outlier data may be because of the fact that the boundaries of the chromosomal aberrations are not known at high resolution.

Two sets of calculations were made from the data. For the first set we used CGH alterations as the independent variable and estimated the frequency of expression alterations in these chromosomal areas. In general, areas with a strong gain of chromosomal material contained a cluster of genes having increased mRNA expression. For example, both chromosomes 1q21-q25, 2p and 9q, showed a relative gain of more than 100% in DNA copy number that was accompanied by increased mRNA expression levels in the two tumor pairs (Fig. 1). In most cases, chromosomal gains detected by CGH were accompanied by an increased level of transcripts in both TCCs 733 (77%) and 827 (80%) (Table I, *top*). Chromosomal losses, on the other hand, were not accompanied by decreased expression in several cases, and were often registered as having unaltered RNA levels (Table I, *top*). The inability to detect RNA expression changes in these cases was not because of fewer genes mapping to the lost regions (data not shown).

In the second set of calculations we selected expression alterations above 2-fold as the independent variable and estimated the frequency of CGH alterations in these areas. As above, we found that increased transcript expression correlated with gain of chromosomal material (TCC 733, 69% and TCC 827, 59%), whereas reduced expression was often detected in areas with unaltered CGH ratios (Table I, *bottom*). Furthermore, as a control we looked at areas with no alter-

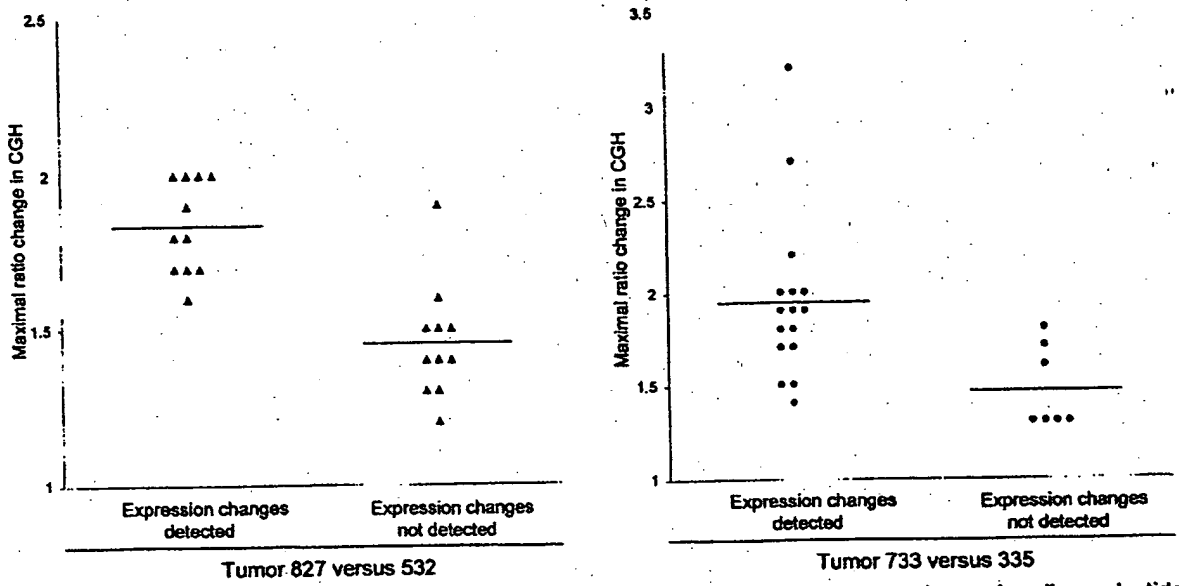


FIG. 2. **Correlation between maximum CGH aberration and the ability to detect expression change by oligonucleotide array monitoring.** The aberration is shown as a numerical -fold change in ratio between invasive tumors 827 (▲) and 733 (♦) and their non-invasive counterparts 532 and 335. The expression change was taken from the *Expression* line to the *right* in Fig. 1, which depicts the resulting expression change for a given chromosomal region. At least half of the mRNAs from a given region have to be either up- or down-regulated to be scored as an expression change. All chromosomal arms in which the CGH ratio plus or minus one standard deviation was outside the ratio value of one were included.

ation in expression. No alteration was detected by CGH in most of these areas (TCC 733, 60% and TCC 827, 81%; see Table I, *bottom*). Because the ability to observe reduced or increased mRNA expression clustering to a certain chromosomal area clearly reflected the extent of copy number changes, we plotted the maximum CGH aberrations in the regions showing CGH changes against the ability to detect a change in mRNA expression as monitored by the oligonucleotide arrays (Fig. 2). For both tumors TCC 733 ($p < 0.015$) and TCC 827 ($p < 0.00003$) a highly significant correlation was observed between the level of CGH ratio change (reflecting the DNA copy number) and alterations detected by the array based technology (Fig. 2). Similar data were obtained when areas with altered expression were used as independent variables. These areas correlated best with CGH when the CGH ratio deviated 1.6- to 2.0-fold (Table I, *bottom*) but mostly did not at lower CGH deviations. These data probably reflect that loss of an allele may only lead to a 50% reduction in expression level, which is at the cut-off point for detection of expression alterations. Gain of chromosomal material can occur to a much larger extent.

Microsatellite-based Detection of Minor Areas of Losses—In TCC 733, several chromosomal areas exhibiting DNA amplification were preceded or followed by areas with a normal CGH but reduced mRNA expression (see Fig. 1, TCC 733 chromosome 1q32, 2p21, and 7q21 and q32, 9q34, and 10q22). To determine whether these results were because of undetected loss of chromosomal material in these regions or because of other non-structural mechanisms regulating transcription, we examined two microsatellites positioned at chromosome 1q25–32 and two at chromosome 2p22. Loss of heterozygosity (LOH) was found at both 1q25 and at 2p22 indicating that minor deleted areas were not detected with the resolution of CGH (Fig. 3). Additionally, chromosome 2p in TCC 733 showed a CGH pattern of gain/no change/gain of DNA that correlated with transcript increase/decrease/increase. Thus, for the areas showing increased expression there was a correlation with the DNA copy number alterations (Fig. 1*A*). As indicated above, the mRNA decrease observed in the middle of the chromosomal gain was because of LOH, implying that one of the mechanisms for mRNA down-regulation may be regions that have undergone smaller losses of chromosomal material. However, this cannot be detected with the resolution of the CGH method.

In both TCC 733 and TCC 827, the telomeric end of chromosome 11p showed a normal ratio in the CGH analysis; however, clusters of five and three genes, respectively, lost their expression. Two microsatellites (D11S1760, D11S922) positioned close to MUC2, IGF2, and cathepsin D indicated LOH as the most likely mechanism behind the loss of expression (data not shown).

A reduced expression of mRNA observed in TCC 733 at chromosomes 3q24, 11p11, 12p12.2, 12q21.1, and 16q24 and in TCC 827 at chromosome 11p15.5, 12p11, 15q11.2, and 18q12 was also examined for chromosomal losses using microsatellites positioned as close as possible to the gene loci

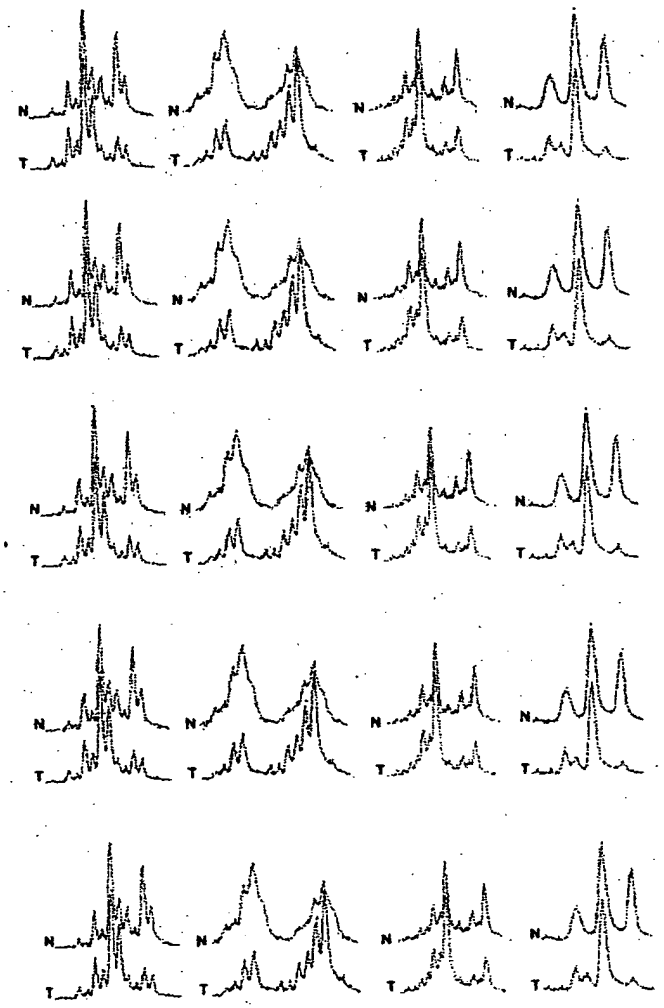


FIG. 3. **Microsatellite analysis of loss of heterozygosity.** Tumor 733 showing loss of heterozygosity at chromosome 1q25, detected (*a*) by D1S215 close to Hu class I histocompatibility antigen (gene number 38 in Fig. 1), (*b*) by D1S2735 close to cathepsin E (gene number 41 in Fig. 1), and (*c*) at chromosome 2p23 by D2S2251 close to general β-spectrin (gene number 11 on Fig. 1) and of (*d*) tumor 827 showing loss of heterozygosity at chromosome 18q12 by S18S1118 close to mitochondrial 3-oxoacyl-coenzyme A thiolase (gene number 12 in Fig. 1). The *upper curves* show the electropherogram obtained from normal DNA from leukocytes (*N*), and the *lower curves* show the electropherogram from tumor DNA (*T*). In all cases one allele is partially lost in the tumor amplicon.

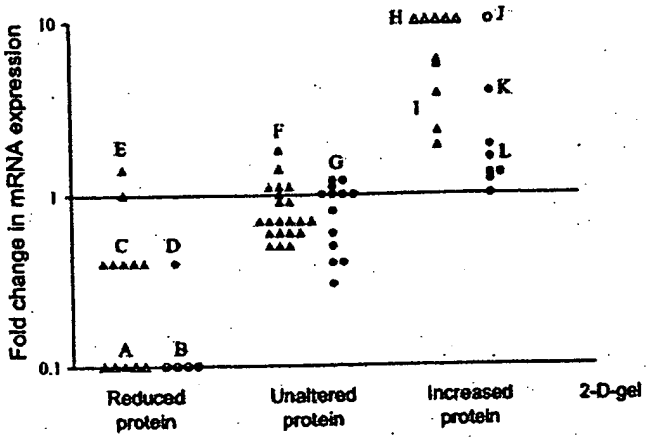


FIG. 4. **Correlation between protein levels as judged by 2D-PAGE and transcript ratio.** For comparison proteins were divided in three groups, unaltered in level or up- or down-regulated (*horizontal axis*). The mRNA ratio as determined by oligonucleotide arrays was plotted for each gene (*vertical axis*). ▲, mRNAs that were scored as present in both tumors used for the ratio calculation; Δ, mRNAs that were scored as absent in the invasive tumors (along *horizontal axis*) or as absent in non-invasive reference (*top* of figure). Two different scalings were used to exclude scaling as a confounder, TCCs 827 and 532 (▲Δ) were scaled with background suppression, and TCCs 733 and 335 (●○) were scaled without suppression. Both comparisons showed highly significant ($p < 0.005$) differences in mRNA ratios between the groups. Proteins shown were as follows: Group *A* (from *left*), phosphoglucomutase 1, glutathione transferase class μ number 4, fatty acid-binding protein homologue, cytokeratin 15, and cytokeratin 13; *B* (from *left*), fatty acid-binding protein homologue, 28-kDa heat shock protein, cytokeratin 13, and calcyclin; *C* (from *left*), α-enolase, hnRNP B1, 28-kDa heat shock protein, 14–3-3-ε, and pre-mRNA splicing factor; *D*, mesothelial keratin K7 (type II); *E* (from *top*), glutathione *S*-transferase-π and mesothelial keratin K7 (type II); *F* (from *top* and *left*), adenylyl cyclase-associated protein, E-cadherin, keratin 19, calgizzarin, phosphoglycerate mutase, annexin IV, cytoskeletal γ-actin, hnRNP A1, integral membrane protein calnexin (IP90), hnRNP H, brain-type clathrin light chain-a, hnRNP F, 70-kDa heat shock protein, heterogeneous nuclear ribonucleoprotein A/B, translationally controlled tumor protein, liver glyceraldehyde-3-phosphate dehydrogenase, keratin 8, aldehyde reductase, and Na,K-ATPase β-1 subunit; *G*, (from *top* and *left*), TCP20, calgizzarin, 70-kDa heat shock protein, calnexin, hnRNP H, cytokeratin 15, ATP synthase, keratin 19, triosephosphate isomerase, hnRNP F, liver glyceraldehyde-3-phosphatase dehydrogenase, glutathione *S*-transferase-π, and keratin 8; *H* (from *left*), plasma gelsolin, autoantigen calreticulin, thioredoxin, and NAD+-dependent 15 hydroxyprostaglandin dehydrogenase; *I* (from *top*), prolyl 4-hydroxylase β-subunit, cytokeratin 20, cytokeratin 17, prohibition, and fructose 1,6-biphosphatase; *J* annexin II; *K*, annexin IV; *L* (from *top* and *left*), 90-kDa heat shock protein, prolyl 4-hydroxylase β-subunit, α-enolase, GRP 78, cyclophilin, and cofilin.

showing reduced mRNA transcripts. Only the microsatellite positioned at 18q12 showed LOH (Fig. 3), suggesting that transcriptional down-regulation of genes in the other regions may be controlled by other mechanisms.

Relation between Changes in mRNA and Protein Levels—2D-PAGE analysis, in combination with Coomassie Brilliant Blue and/or silver staining, was carried out on all four tumors using fresh biopsy material. 40 well resolved abundant known proteins migrating in areas away from the edges of the pH gradient, and having a known chromosomal location, were selected for analysis in the TCC pair 827/532. Proteins were identified by a combination of methods (see "Experimental Procedures"). In general there was a highly significant correlation ($p < 0.005$) between mRNA and protein alterations (Fig. 4). Only one gene showed disagreement between transcript alteration and protein alteration. Except for a group of cyto-

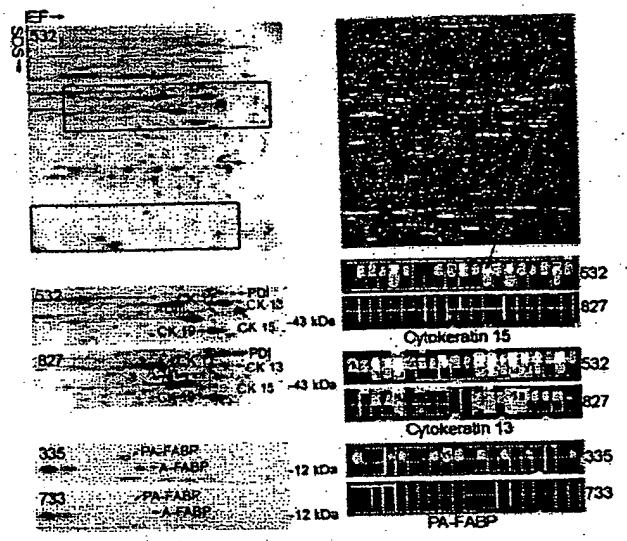


FIG. 5. **Comparison of protein and transcript levels in invasive and non-invasive TCCs.** The *upper part* of the figure shows a 2D gel (*left*) and the oligonucleotide array (*right*) of TCC 532. The *red rectangles* on the *upper gel* highlight the areas that are compared below. Identical areas of 2D gels of TCCs 532 and 827 are shown below. Clearly, cytokeratins 13 and 15 are strongly down-regulated in TCC 827 (*red annotation*). The tile on the array containing probes for cytokeratin 15 is enlarged *below* the array (*red arrow*) from TCC 532 and is compared with TCC 827. The *upper row* of *squares* in each tile corresponds to perfect match probes; the *lower row* corresponds to mismatch probes containing a mutation (used for correction for unspecific binding). Absence of signal is depicted as *black*, and the higher the signal the lighter the color. A high transcript level was detected in TCC 532 (6151 units) whereas a much lower level was detected in TCC 827 (absence of signals). For cytokeratin 13, a high transcript level was also present in TCC 532 (15659 units), and a much lower level was present in TCC 827 (623 units). The 2D gels at the *bottom* of the figure (*left*) show levels of PA-FABP and adipocyte-FABP in TCCs 335 and 733 (invasive), respectively. Both proteins are down-regulated in the invasive tumor. To the *right* we show the array tiles for the PA-FABP transcript. A medium transcript level was detected in the case of TCC 335 (1277 units) whereas very low levels were detected in TCC 733 (166 units). *IEF*, isoelectric focusing.

keratins encoded by genes on chromosome 17 (Fig. 5) the analyzed proteins did not belong to a particular family. 26 well focused proteins whose genes had a know chromosomal location were detected in TCCs 733 and 335, and of these 19 correlated ($p < 0.005$) with the mRNA changes detected using the arrays (Fig. 4). For example, PA-FABP was highly expressed in the non-invasive TCC 335 but lost in the invasive counterpart (TCC 733; see Fig. 5). The smaller number of proteins detected in both 733 and 335 was because of the smaller size of the biopsies that were available.

11 chromosomal regions where CGH showed aberrations that corresponded to the changes in transcript levels also showed corresponding changes in the protein level (Table II). These regions included genes that encode proteins that are found to be frequently altered in bladder cancer, namely cytokeratins 17 and 20, annexins II and IV, and the fatty acid-binding proteins PA-FABP and FBP1. Four of these proteins were encoded by genes in chromosome 17q, a frequently amplified chromosomal area in invasive bladder cancers.

DISCUSSION

Most human cancers have abnormal DNA content, having lost some chromosomal parts and gained others. The present study provides some evidence as to the effect of these gains and losses on gene expression in two pairs of non-invasive and invasive TCCs using high throughput expression arrays and proteomics, in combination with CGH. In general, the results showed that there is a clear individual regulation of the mRNA expression of single genes, which in some cases was superimposed by a DNA copy number effect. In most cases, genes located in chromosomal areas with gains often exhibited increased mRNA expression, whereas areas showing losses showed either no change or a reduced mRNA expression. The latter might be because of the fact that losses most often are restricted to loss of one allele, and the cut-off point for detection of expression alterations was a 2-fold change, thus being at the border of detection. In several cases, how-

TABLE II

Proteins whose expression level correlates with both mRNA and gene dose changes

Protein	Chromosomal location	Tumor TCC	CGH alteration	Transcript alteration[a]	Protein alteration
Annexin II	1q21	733	Gain	Abs to Pres[a]	Increase
Annexin IV	2p13	733	Gain	3.9-Fold up	Increase
Cytokeratin 17	17q12-q21	827	Gain	3.8-Fold up	Increase
Cytokeratin 20	17q21.1	827	Gain	5.6-Fold up	Increase
(PA-)FABP	8q21.2	827	Loss	10-Fold down	Decrease
FBP1	9q22	827	Gain	2.3-Fold up	Increase
Plasma gelsolin	9q31	827	Gain	Abs to Pres	Increase
Heat shock protein 28	15q12-q13	827	Loss	2.5-Fold up	Decrease
Prohibitin	17q21	827/733	Gain	3.7-/2.5-Fold up[b]	Increase
Prolyl-4-hydroxyl	17q25	827/733	Gain	5.7-/1.6-Fold up	Increase
hnRNPB1	7p15	827	Loss	2.5-Fold down	Decrease

[a] Abs, absent; Pres, present.
[b] In cases where the corresponding alterations were found in both TCCs 827 and 733 these are shown as 827/733.

ever, an increase or decrease in DNA copy number was associated with *de novo* occurrence or complete loss of transcript, respectively. Some of these transcripts could not be detected in the non-invasive tumor but were present at relatively high levels in areas with DNA amplifications in the invasive tumors (*e.g.* in TCC 733 transcript from cellular ligand of annexin II gene (chromosome 1q21) from absent to 2670 arbitrary units; in TCC 827 transcript from small proline-rich protein 1 gene (chromosome 1q12-q21.1) from absent to 1326 arbitrary units). It may be anticipated from these data that significant clustering of genes with an increased expression to a certain chromosomal area indicates an increased likelihood of gain of chromosomal material in this area.

Considering the many possible regulatory mechanisms acting at the level of transcription, it seems striking that the gene dose effects were so clearly detectable in gained areas. One hypothetical explanation may lie in the loss of controlled methylation in tumor cells (17–19). Thus, it may be possible that in chromosomes with increased DNA copy numbers two or more alleles could be demethylated simultaneously leading to a higher transcription level, whereas in chromosomes with losses the remaining allele could be partly methylated, turning off the process (20, 21). A recent report has documented a ploidy regulation of gene expression in yeast, but in this case all the genes were present in the same ratio (22), a situation that is not analogous to that of cancer cells, which show marked chromosomal aberrations, as well as gene dosage effects.

Several CGH studies of bladder cancer have shown that some chromosomal aberrations are common at certain stages of disease progression, often occurring in more than 1 of 3 tumors. In pTa tumors, these include 9p−, 9q−, 1q+, Y− (2, 6), and in pT1 tumors, 2q−,11p−, 11q−, 1q+, 5p+, 8q+, 17q+, and 20q+ (2–4, 6, 7). The pTa tumors studied here showed similar aberrations such as 9p− and 9q22-q33− and 9q− and Y−, respectively. Likewise, the two minimal invasive pT1 tumors showed aberrations that are commonly seen at that stage, and TCC 827 had a remarkable resemblance to the commonly seen pattern of losses and gains, such as 1q22–24 amplification (seen in both tumors), 11q14-q22 loss, the latter often linked to 17 q+ (both tumors), and 1q+ and 9p−, often linked to 20q+ and 11 q13+ (both tumors) (7–9). These observations indicate that the pairs of tumors used in this study exhibit chromosomal changes observed in many tumors, and therefore the findings could be of general importance for bladder cancer.

Considering that the mapping resolution of CGH is of about 20 megabases it is only possible to get a crude picture of chromosomal instability using this technique. Occasionally, we observed reduced transcript levels close to or inside regions with increased copy numbers. Analysis of these regions by positioning heterozygous microsatellites as close as possible to the locus showing reduced gene expression revealed loss of heterozygosity in several cases. It seems likely that multiple and different events occur along each chromosomal arm and that the use of cDNA microarrays for analysis of DNA copy number changes will reach a resolution that can resolve these changes, as has recently been proposed (2). The outlier data were not more frequent at the boundaries of the CGH aberrations. At present we do not know the mechanism behind chromosomal aneuploidy and cannot predict whether chromosomal gains will be transcribed to a larger extent than the two native alleles. A mechanism as genetic imprinting has an impact on the expression level in normal cells and is often reduced in tumors. However, the relation between imprinting and gain of chromosomal material is not known.

We regard it as a strength of this investigation that we were able to compare invasive tumors to benign tumors rather than to normal urothelium, as the tumors studied were biologically very close and probably may represent successive steps in the progression of bladder cancer. Despite the limited amount of fresh tissue available it was possible to apply three different state of the art methods. The observed correlation between DNA copy number and mRNA expression is remarkable when one considers that different pieces of the tumor biopsies were used for the different sets of experiments. This indicate that bladder tumors are relatively homogenous, a notion recently supported by CGH and LOH data that showed a remarkable similarity even between tumors and distant metastasis (10, 23).

In the few cases analyzed, mRNA and protein levels showed a striking correspondence although in some cases we found discrepancies that may be attributed to translational regulation, post-translational processing, protein degradation, or a combination of these. Some transcripts belong to undertranslated mRNA pools, which are associated with few translationally inactive ribosomes; these pools, however, seem to be rare (24). Protein degradation, for example, may be very important in the case of polypeptides with a short half-life (*e.g.* signaling proteins). A poor correlation between mRNA and protein levels was found in liver cells as determined by arrays and 2D-PAGE (25), and a moderate correlation was recently reported by Ideker *et al.* (26) in yeast.

Interestingly, our study revealed a much better correlation between gained chromosomal areas and increased mRNA levels than between loss of chromosomal areas and reduced mRNA levels. In general, the level of CGH change determined the ability to detect a change in transcript. One possible explanation could be that by losing one allele the change in mRNA level is not so dramatic as compared with gain of material, which can be rather unlimited and may lead to a severalfold increase in gene copy number resulting in a much higher impact on transcript level. The latter would be much easier to detect on the expression arrays as the cut-off point was placed at a 2-fold level so as not to be biased by noise on the array. Construction of arrays with a better signal to noise ratio may in the future allow detection of lesser than 2-fold alterations in transcript levels, a feature that may facilitate the analysis of the effect of loss of chromosomal areas on transcript levels.

In eleven cases we found a significant correlation between DNA copy number, mRNA expression, and protein level. Four of these proteins were encoded by genes located at a frequently amplified area in chromosome 17q. Whether DNA copy number is one of the mechanisms behind alteration of these eleven proteins is at present unknown and will have to be proved by other methods using a larger number of samples. One factor making such studies complicated is the large extent of protein modification that occurs after translation, requiring immunoidentification and/or mass spectrometry to correctly identify the proteins in the gels.

In conclusion, the results presented in this study exemplify the large body of knowledge that may be possible to gather in the future by combining state of the art techniques that follow the pathway from DNA to protein (26). Here, we used a traditional chromosomal CGH method, but in the future high resolution CGH based on microarrays with many thousand radiation hybrid-mapped genes will increase the resolution and information derived from these types of experiments (2). Combined with expression arrays analyzing transcripts derived from genes with known locations, and 2D gel analysis to obtain information at the post-translational level, a clearer and more developed understanding of the tumor genome will be forthcoming.

Acknowledgments—We thank Mie Madsen, Hanne Steen, Inge Lis Thorsen, Hans Lund, Vikolaj Ørntoft, and Lynn Bjerke for technical help and Thomas Gingeras, Christine Harrington, and Morten Østergaard for valuable discussions.

* This work was supported by grants from The Danish Cancer Society, the University of Aarhus, Aarhus County, Novo Nordic, the Danish Biotechnology Program, the Frenkels Foundation, the John and Birthe Meyer Foundation, and NCI, National Institutes of Health Grant CA47537. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: Dept. of Clinical Biochemistry, Molecular Diagnostic Laboratory, Aarhus University Hospital, Skejby, DK-8200 Aarhus N, Denmark. Tel.: 45-89495100/45-86156201 (private); Fax: 45-89496018; E-mail: orntoft@kba.sks.au.dk.

REFERENCES

1. Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998) Genetic instabilities in human cancers. *Nature* **17**, 643–649
2. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**, 41–46
3. de Cremoux, P., Martin, E. C., Vincent-Salomon, A., Dieras, V., Barbaroux, C., Liva, S., Pouillart, P., Sastre-Garau, X., and Magdelenat, H. (1999) Quantitative PCR analysis of c-erb B-2 (HER2/neu) gene amplification and comparison with p185(HER2/neu) protein expression in breast cancer drill biopsies. *Int. J. Cancer* **83**, 157–161
4. Brungier, P. P., Tamimi, Y., Shuuring, E., and Schalken, J. (1996) Expression of cyclin D1 and EMS1 in bladder tumors; relationship with chromosome 11q13 amplifications. *Oncogene* **12**, 1747–1753
5. Slavc, I., Ellenbogen, R., Jung, W. H., Vawter, G. F., Kretschmar, C., Grier, H., and Korf, B. R. (1990) *myc* gene amplification and expression in primary human neuroblastoma. *Cancer Res.* **50**, 1459–1463
6. Sauter, G., Carroll, P., Moch, H., Kallioniemi, A., Kerschmann, R., Narayan, P., Mihatsch, M. J., and Waldman, F. M. (1995) c-*myc* copy number gains in bladder cancer detected by fluorescence *in situ* hybridization. *Am J. Pathol.* **146**, 1131–1139
7. Richter, J., Jiang, F., Gorog, J. P., Sartorious, G., Egenter, C., Gasser, T. C., Moch, H., Mihatsch, M. J., and Sauter, G. (1997) Marked genetic differences between stage pTa and stage pT1 papillary bladder cancer detected by comparative genomic hybridization. *Cancer Res.* **57**, 2860–2864
8. Richter, J., Beffa, L., Wagner, U., Schraml, P., Gasser, T. C., Moch, H., Mihatsch, M. J., and Sauter, G. (1998) Patterns of chromosomal imbalances in advanced urinary bladder cancer detected by comparative genomic hybridization. *Am. J. Pathol.* **153**, 1615–1621
9. Bruch, J., Wohr, G., Hautmann, R., Mattfeldt, T., Bruderlein, S., Moller, P., Sauter, S., Hameister, H., Vogel, W., and Paiss, T. (1998) Chromosomal changes during progression of transitional cell carcinoma of the bladder and delineation of the amplified interval on chromosome arm 8q. *Genes Chromosomes Cancer* **23**, 167–174
10. Hovey, R. M., Chu, L., Balazs, M., De Vries, S., Moore, D., Sauter, G., Carroll, P. R., and Waldman, F. M. (1998) Genetic alterations in primary bladder cancers and their metastases. *Cancer Res.* **15**, 3555–3560
11. Simon, R., Burger, H., Brinkschmidt, C., Bocker, W., Hertle, L., and Terpe, H. J. (1998) Chromosomal aberrations associated with invasion in papillary superficial bladder cancer. *J. Pathol.* **185**, 345–351
12. Koo, S. H., Kwon, K. C., Ihm, C. H., Jeon, Y. M., Park, J. W., and Sul, C. K. (1999) Detection of genetic alterations in bladder tumors by comparative genomic hybridization and cytogenetic analysis. *Cancer Genet. Cytogenet.* **110**, 87–93
13. Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**, 1359–1367
14. Christensen, M., Sunde, L., Bolund, L., and Ornfoft, T. F. (1999) Comparison of three methods of microsatellite detection. *Scand. J. Clin. Lab. Invest.* **59**, 167–177
15. Celis, J. E., Ostergaard, M., Basse, B., Celis, A., Lauridsen, J. B., Ratz, G. P., Andersen, I., Hein, B., Wolf, H., Ornfoft, T. F., and Rasmussen, H. H. (1996) Loss of adipocyte-type fatty acid binding protein and other protein biomarkers is associated with progression of human bladder transitional cell carcinomas. *Cancer Res.* **56**, 4782–4790
16. Celis, J. E., Ratz, G., Basse, B., Lauridsen, J. B., and Celis, A. (1994) in *Cell Biology: A Laboratory Handbook* (Celis, J. E., ed) Vol. 3, pp. 222–230, Academic Press, Orlando, FL
17. Ohlsson, R., Tycko, B., and Sapienza, C. (1998) Monoallelic expression: 'there can only be one'. *Trends Genet.* **14**, 435–438
18. Hollander, G. A., Zuklys, S., Morel, C., Mizoguchi, E., Mobisson, K., Simpson, S., Terhorst, C., Wishart, W., Golan, D. E., Bhan, A. K., and Burakoff, S. J. (1998) Monoallelic expression of the interleukin-2 locus. *Science* **279**, 2118–2121
19. Brannan, C. I., and Bartolomei, M. S. (1999) Mechanisms of genomic imprinting. *Curr. Opin. Genet. Dev.* **9**, 164–170
20. Ohlsson, R., Cui, H., He, L., Pfeifer, S., Malmikumpa, H., Jiang, S., Feinberg, A. P., and Hedborg, F. (1999) Mosaic allelic insulin-like growth factor 2 expression patterns reveal a link between Wilms' tumorigenesis and epigenetic heterogeneity. *Cancer Res.* **59**, 3889–3892
21. Cui, H., Hedborg, F., He, L., Nordenskjold, A., Sandstedt, B., Pfeifer-Ohlsson, S., and Ohlsson, R. (1997) Inactivation of H19, an imprinted and putative tumor repressor gene, is a preneoplastic event during Wilms' tumorigenesis. *Cancer Res.* **57**, 4469–4473
22. Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S., and Fink, G. R. (1999) Ploidy regulation of gene expression. *Science* **285**, 251–254
23. Tsao, J., Yatabe, Y., Markl, I. D., Hajyan, K., Jones, P. A., and Shibata, D. (2000) Bladder cancer genotype stability during clinical progression. *Genes Chromosomes Cancer* **29**, 26–32
24. Zong, Q., Schummer, M., Hood, L., and Morris, D. R. (1999) Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10632–10636
25. Anderson, L., and Seilhamer, J. (1997) Comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533–537
26. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934

[CANCER RESEARCH 62, 6240–6245, November 1, 2002]

Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer[1,2]

Elizabeth Hyman,[3] Päivikki Kauraniemi,[3] Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahloun, Olli-P. Kallioniemi, and Anne Kallioniemi[4]

Howard Hughes Medical Institute-NIH Research Scholar, Bethesda, Maryland 20892 [E. H.]; Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland 20892 [E. H., P. K., S. H., M. W., S. M., E. R., M. R., A. E., O. K., A. K.]; Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, FIN-33520 Tampere, Finland [P. K., A. K.]; Signal Processing Laboratory, Tampere University of Technology, FIN-33101 Tampere, Finland [S. H.]; Institute of Pathology, University of Basel, CH-4003 Basel, Switzerland [G. S.]; and Biomedicum Biochip Center, Helsinki University Hospital, Biomedicum Helsinki, FIN-00014 Helsinki, Finland [O. M.]

ABSTRACT

Genetic changes underlie tumor progression and may lead to cancer-specific expression of critical genes. Over 1100 publications have described the use of comparative genomic hybridization (CGH) to analyze the pattern of copy number alterations in cancer, but very few of the genes affected are known. Here, we performed high-resolution CGH analysis on cDNA microarrays in breast cancer and directly compared copy number and mRNA expression levels of 13,824 genes to quantitate the impact of genomic changes on gene expression. We identified and mapped the boundaries of 24 independent amplicons, ranging in size from 0.2 to 12 Mb. Throughout the genome, both high- and low-level copy number changes had a substantial impact on gene expression, with 44% of the highly amplified genes showing overexpression and 10.5% of the highly overexpressed genes being amplified. Statistical analysis with random permutation tests identified 270 genes whose expression levels across 14 samples were systematically attributable to gene amplification. These included most previously described amplified genes in breast cancer and many novel targets for genomic alterations, including the *HOXB7* gene, the presence of which in a novel amplicon at 17q21.3 was validated in 10.2% of primary breast cancers and associated with poor patient prognosis. In conclusion, CGH on cDNA microarrays revealed hundreds of novel genes whose overexpression is attributable to gene amplification. These genes may provide insights to the clonal evolution and progression of breast cancer and highlight promising therapeutic targets.

INTRODUCTION

Gene expression patterns revealed by cDNA microarrays have facilitated classification of cancers into biologically distinct categories, some of which may explain the clinical behavior of the tumors (1–6). Despite this progress in diagnostic classification, the molecular mechanisms underlying gene expression patterns in cancer have remained elusive, and the utility of gene expression profiling in the identification of specific therapeutic targets remains limited.

Accumulation of genetic defects is thought to underlie the clonal evolution of cancer. Identification of the genes that mediate the effects of genetic changes may be important by highlighting transcripts that are actively involved in tumor progression. Such transcripts and their encoded proteins would be ideal targets for anticancer therapies, as demonstrated by the clinical success of new therapies against amplified oncogenes, such as *ERBB2* and *EGFR* (7, 8), in breast cancer and other solid tumors. Besides amplifications of known oncogenes, over 20 recurrent regions of DNA amplification have been mapped in breast cancer by CGH[5] (9, 10). However, these amplicons are often large and poorly defined, and their impact on gene expression remains unknown.

We hypothesized that genome-wide identification of those gene expression changes that are attributable to underlying gene copy number alterations would highlight transcripts that are actively involved in the causation or maintenance of the malignant phenotype. To identify such transcripts, we applied a combination of cDNA and CGH microarrays to: (*a*) determine the global impact that gene copy number variation plays in breast cancer development and progression; and (*b*) identify and characterize those genes whose mRNA expres-

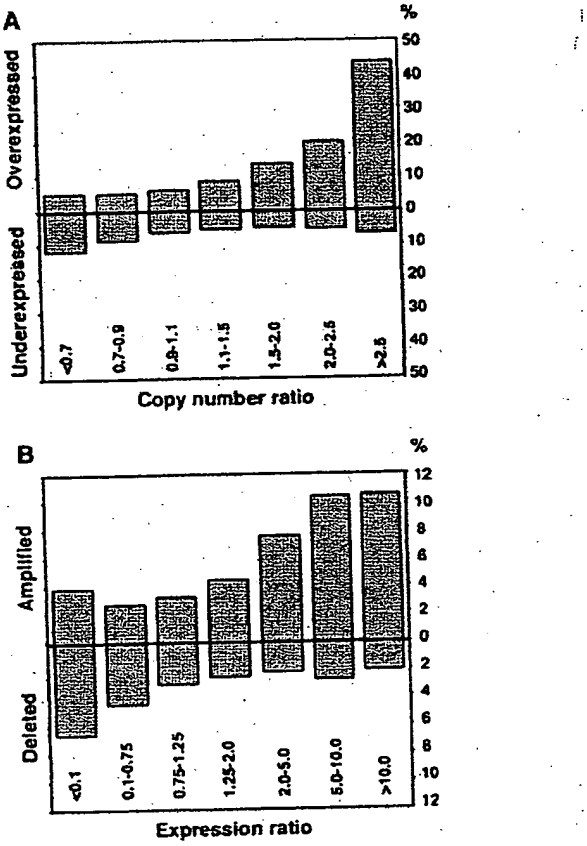


Fig. 1. Impact of gene copy number on global gene expression levels. *A*, percentage of over- and underexpressed genes (*Y axis*) according to copy number ratios (*X axis*). Threshold values used for over- and underexpression were >2.184 (global upper 7% of the cDNA ratios) and <0.4826 (global lower 7% of the expression ratios). *B*, percentage of amplified and deleted genes according to expression ratios. Threshold values for amplification and deletion were >1.5 and <0.7.

Received 5/29/02; accepted 8/28/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported in part by the Academy of Finland, Emil Aaltonen Foundation, the Finnish Cancer Society, the Pirkanmaa Cancer Society, the Pirkanmaa Cultural Foundation, the Finnish Breast Cancer Group, the Foundation for the Development of Laboratory Medicine, the Medical Research Fund of the Tampere University Hospital, the Foundation for Commercial and Technical Sciences, and the Swedish Research Council.

[2] Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org).

[3] Contributed equally to this work.

[4] To whom requests for reprints should be addressed, at Laboratory of Cancer Genetics, Institute of Medical Technology, Lenkkeilijankatu 6, FIN-33520 Tampere, Finland. Phone: 358-3247-4125; Fax: 358-3247-4168; E-mail: anne.kallioniemi@uta.fi.

[5] The abbreviations used are: CGH, comparative genomic hybridization; FISH, fluorescence *in situ* hybridization; RT-PCR, reverse transcription-PCR.

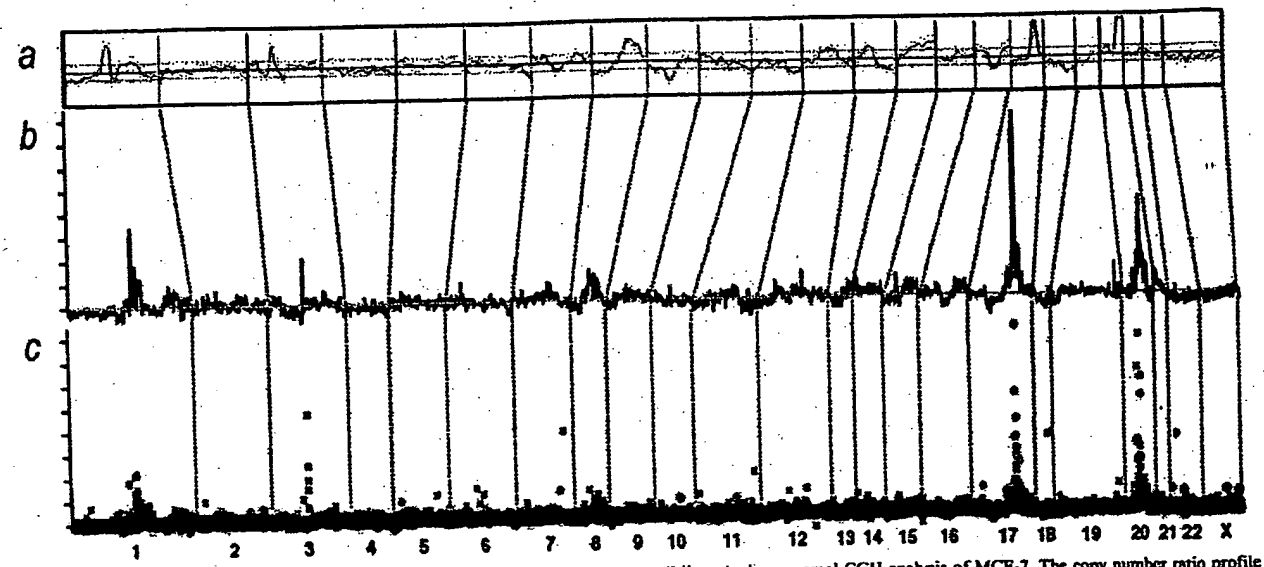


Fig. 2. Genome-wide copy number and expression analysis in the MCF-7 breast cancer cell line. *A*, chromosomal CGH analysis of MCF-7. The copy number ratio profile (*blue line*) across the entire genome from 1p telomere to Xq telomere is shown along with ±1 SD (*orange lines*). The *black horizontal line* indicates a ratio of 1.0; *red line*, a ratio of 0.8; and *green line*, a ratio of 1.2. *B–C*, genome-wide copy number analysis in MCF-7 by CGH on cDNA microarray. The copy number ratios were plotted as a function of the position of the cDNA clones along the human genome. In *B*, individual data points are connected with a line, and a moving median of 10 adjacent clones is shown. *Red horizontal line*, the copy number ratio of 1.0. In *C*, individual data points are labeled by color coding according to cDNA expression ratios. The *bright red dots* indicate the upper 2%, and *dark red dots*, the next 5% of the expression ratios in MCF-7 cells (overexpressed genes); *bright green dots* indicate the lowest 2%, and *dark green dots*, the next 5% of the expression ratios (underexpressed genes); the rest of the observations are shown with *black crosses*. The chromosome numbers are shown at the *bottom* of the figure, and chromosome boundaries are indicated with a *dashed line*.

sion is most significantly associated with amplification of the corresponding genomic template.

MATERIALS AND METHODS

Breast Cancer Cell Lines. Fourteen breast cancer cell lines (BT-20, BT-474, HCC1428, Hs578t, MCF7, MDA-361, MDA-436, MDA-453, MDA-468, SKBR-3, T-47D, UACC812, ZR-75-1, and ZR-75-30) were obtained from the American Type Culture Collection (Manassas, VA). Cells were grown under recommended culture conditions. Genomic DNA and mRNA were isolated using standard protocols.

Copy Number and Expression Analyses by cDNA Microarrays. The preparation and printing of the 13,824 cDNA clones on glass slides were performed as described (11–13). Of these clones, 244 represented uncharacterized expressed sequence tags, and the remainder corresponded to known genes. CGH experiments on cDNA microarrays were done as described (14, 15). Briefly, 20 μg of genomic DNA from breast cancer cell lines and normal human WBCs were digested for 14–18 h with *Alu*I and *Rsa*I (Life Technologies, Inc., Rockville, MD) and purified by phenol/chloroform extraction. Six μg of digested cell line DNAs were labeled with Cy3-dUTP (Amersham Pharmacia) and normal DNA with Cy5-dUTP (Amersham Pharmacia) using the Bioprime Labeling kit (Life Technologies, Inc.). Hybridization (14, 15) and posthybridization washes (13) were done as described. For the expression analyses, a standard reference (Universal Human Reference RNA; Stratagene, La Jolla, CA) was used in all experiments. Forty μg of reference RNA were labeled with Cy3-dUTP and 3.5 μg of test mRNA with Cy5-dUTP, and the labeled cDNAs were hybridized on microarrays as described (13, 15). For both microarray analyses, a laser confocal scanner (Agilent Technologies, Palo Alto, CA) was used to measure the fluorescence intensities at the target locations using the DEARRAY software (16). After background subtraction, average intensities at each clone in the test hybridization were divided by the average intensity of the corresponding clone in the control hybridization. For the copy number analysis, the ratios were normalized on the basis of the distribution of ratios of all targets on the array and for the expression analysis on the basis of 88 housekeeping genes, which were spotted four times onto the array. Low quality measurements (*i.e.*, copy number data with mean reference intensity <100 fluorescent units, and expression data with both test and reference intensity <100 fluorescent units and/or with spot size <50 units) were excluded from the analysis and were treated as missing values. The distributions of fluorescence ratios were used to define cutpoints for increased/decreased copy number. Genes with CGH ratio >1.43 (representing the upper 5% of the CGH ratios across all experiments) were considered to be amplified, and genes with ratio <0.73 (representing the lower 5%) were considered to be deleted.

Statistical Analysis of CGH and cDNA Microarray Data. To evaluate the influence of copy number alterations on gene expression, we applied the following statistical approach. CGH and cDNA calibrated intensity ratios were log-transformed and normalized using median centering of the values in each cell line. Furthermore, cDNA ratios for each gene across all 14 cell lines were median centered. For each gene, the CGH data were represented by a vector that was labeled 1 for amplification (ratio, >1.43) and 0 for no amplification. Amplification was correlated with gene expression using the signal-to-noise statistics (1). We calculated a weight, w_g, for each gene as follows:

$$w_g = \frac{m_{g1} - m_{g0}}{\sigma_{g1} + \sigma_{g0}}$$

where m_{g1}, σ_{g1} and m_{g0}, σ_{g0} denote the means and SDs for the expression levels for amplified and nonamplified cell lines, respectively. To assess the statistical significance of each weight, we performed 10,000 random permutations of the label vector. The probability that a gene had a larger or equal weight by random permutation than the original weight was denoted by α. A low α (<0.05) indicates a strong association between gene expression and amplification.

Genomic Localization of cDNA Clones and Amplicon Mapping. Each cDNA clone on the microarray was assigned to a Unigene cluster using the Unigene Build 141.[6] A database of genomic sequence alignment information for mRNA sequences was created from the August 2001 freeze of the University of California Santa Cruz's GoldenPath database.[7] The chromosome and bp positions for each cDNA clone were then retrieved by relating these data sets. Amplicons were defined as a CGH copy number ratio >2.0 in at least two adjacent clones in two or more cell lines or a CGH ratio >2.0 in at least three adjacent clones in a single cell line. The amplicon start and end positions were

[6] Internet address: http://research.nhgri.nih.gov/microarray/downloadable_cdna.html.
[7] Internet address: www.genome.ucsc.edu.

Table 1 *Summary of independent amplicons in 14 breast cancer cell lines by CGH microarray*

Location	Start (Mb)	End (Mb)	Size (Mb)
1p13	132.79	132.94	0.2
1q21	173.92	177.25	3.3
1q22	179.28	179.57	0.3
3p14	71.94	74.66	2.7
7p12.1–7p11.2	55.62	60.95	5.3
7q31	125.73	130.96	5.2
7q32	140.01	140.68	0.7
8q21.11–8q21.13	86.45	92.46	6.0
8q21.3	98.45	103.05	4.6
8q23.3–8q24.14	129.88	142.15	12.3
8q24.22	151.21	152.16	1.0
9p13	38.65	39.25	0.6
13q22–q31	77.15	81.38	4.2
16q22	86.70	87.62	0.9
17q11	29.30	30.85	1.6
17q12–q21.2	39.79	42.80	3.0
17q21.32–q21.33	52.47	55.80	3.3
17q22–q23.3	63.81	69.70	5.9
17q23.3–q24.3	69.93	74.99	5.1
19q13	40.63	41.40	0.8
20q11.22	34.59	35.85	1.3
20q13.12	44.00	45.62	1.6
20q13.12–q13.13	46.45	49.43	3.0
20q13.2–q13.32	51.32	59.12	7.8

extended to include neighboring nonamplified clones (ratio, <1.5). The amplicon size determination was partially dependent on local clone density.

FISH. Dual-color interphase FISH to breast cancer cell lines was done as described (17). Bacterial artificial chromosome clone RP11-361K8 was labeled with SpectrumOrange (Vysis, Downers Grove, IL), and SpectrumOrange-labeled probe for *EGFR* was obtained from Vysis. SpectrumGreen-labeled chromosome 7 and 17 centromere probes (Vysis) were used as a reference. A tissue microarray containing 612 formalin-fixed, paraffin-embedded primary breast cancers (17) was applied in FISH analyses as described (18). The use of these specimens was approved by the Ethics Committee of the University of Basel and by the NIH. Specimens containing a 2-fold or higher increase in the number of test probe signals, as compared with corresponding centromere signals, in at least 10% of the tumor cells were considered to be amplified. Survival analysis was performed using the Kaplan-Meier method and the log-rank test.

RT-PCR. The *HOXB7* expression level was determined relative to *GAPDH*. Reverse transcription and PCR amplification were performed using Access RT-PCR System (Promega Corp., Madison, WI) with 10 ng of mRNA as a template. *HOXB7* primers were 5′-GAGCAGAGGGACTCGGACTT-3′ and 5′-GCGTCAGGTAGCGATTGTAG-3′.

RESULTS

Global Effect of Copy Number on Gene Expression. 13,824 arrayed cDNA clones were applied for analysis of gene expression and gene copy number (CGH microarrays) in 14 breast cancer cell lines. The results illustrate a considerable influence of copy number on gene expression patterns. Up to 44% of the highly amplified transcripts (CGH ratio, >2.5) were overexpressed (*i.e.*, belonged to the global upper 7% of expression ratios), compared with only 6% for genes with normal copy number levels (Fig. 1A). Conversely, 10.5% of the transcripts with high-level expression (cDNA ratio, >10) showed increased copy number (Fig. 1*B*). Low-level copy number increases and decreases were also associated with similar, although less dramatic, outcomes on gene expression (Fig. 1).

Identification of Distinct Breast Cancer Amplicons. Base-pair locations obtained for 11,994 cDNAs (86.8%) were used to plot copy number changes as a function of genomic position (Fig. 2, Supplement Fig. A). The average spacing of clones throughout the genome was 267 kb. This high-resolution mapping identified 24 independent breast cancer amplicons, spanning from 0.2 to 12 Mb of DNA (Table 1). Several amplification sites detected previously by chromosomal CGH were validated, with 1q21, 17q12–q21.2, 17q22–q23, 20q13.1, and 20q13.2 regions being most commonly amplified. Furthermore, the boundaries of these amplicons were precisely delineated. In addition, novel amplicons were identified at 9p13 (38.65–39.25 Mb), and 17q21.3 (52.47–55.80 Mb).

Direct Identification of Putative Amplification Target Genes. The cDNA/CGH microarray technique enables the direct correlation of copy number and expression data on a gene-by-gene basis throughout the genome. We directly annotated high-resolution CGH plots with gene expression data using color coding. Fig. 2*C* shows that most of the amplified genes in the MCF-7 breast cancer cell line at 1p13, 17q22–q23, and 20q13 were highly overexpressed. A view of chromosome 7 in the MDA-468 cell line implicates *EGFR* as the most highly overexpressed and amplified gene at 7p11–p12 (Fig. 3*A*). In BT-474, the two known amplicons at 17q12 and 17q22–q23 contained numerous highly overexpressed genes (Fig. 3*B*). In addition, several genes, including the homeobox genes *HOXB2* and *HOXB7*, were highly amplified in a previously undescribed independent amplicon at 17q21.3. *HOXB7* was systematically amplified (as validated by FISH, Fig. 3*B*, *inset*) as well as overexpressed (as verified by RT-PCR, data not shown) in BT-474, UACC812, and ZR-75-30 cells. Furthermore, this novel

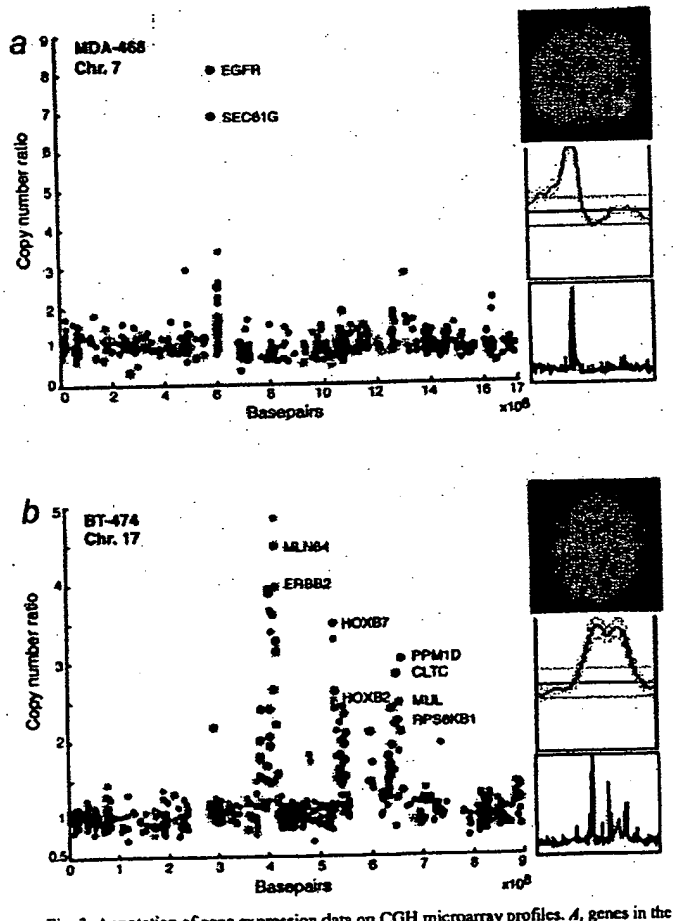


Fig. 3. Annotation of gene expression data on CGH microarray profiles. *A*, genes in the 7p11-p12 amplicon in the MDA-468 cell line are highly expressed (*red dots*) and include the *EGFR* oncogene. *B*, several genes in the 17q12, 17q21.3, and 17q23 amplicons in the BT-474 breast cancer cell line are highly overexpressed (*red*) and include the *HOXB7* gene. The data labels and color coding are as indicated for Fig. 2*C*. *Insets* show chromosomal CGH profiles for the corresponding chromosomes and validation of the increased copy number by interphase FISH using *EGFR* (*red*) and chromosome 7 centromere probe (*green*) to MDA-468 (*A*) and *HOXB7*-specific probe (*red*) and chromosome 17 centromere (*green*) to BT-474 cells (*B*).

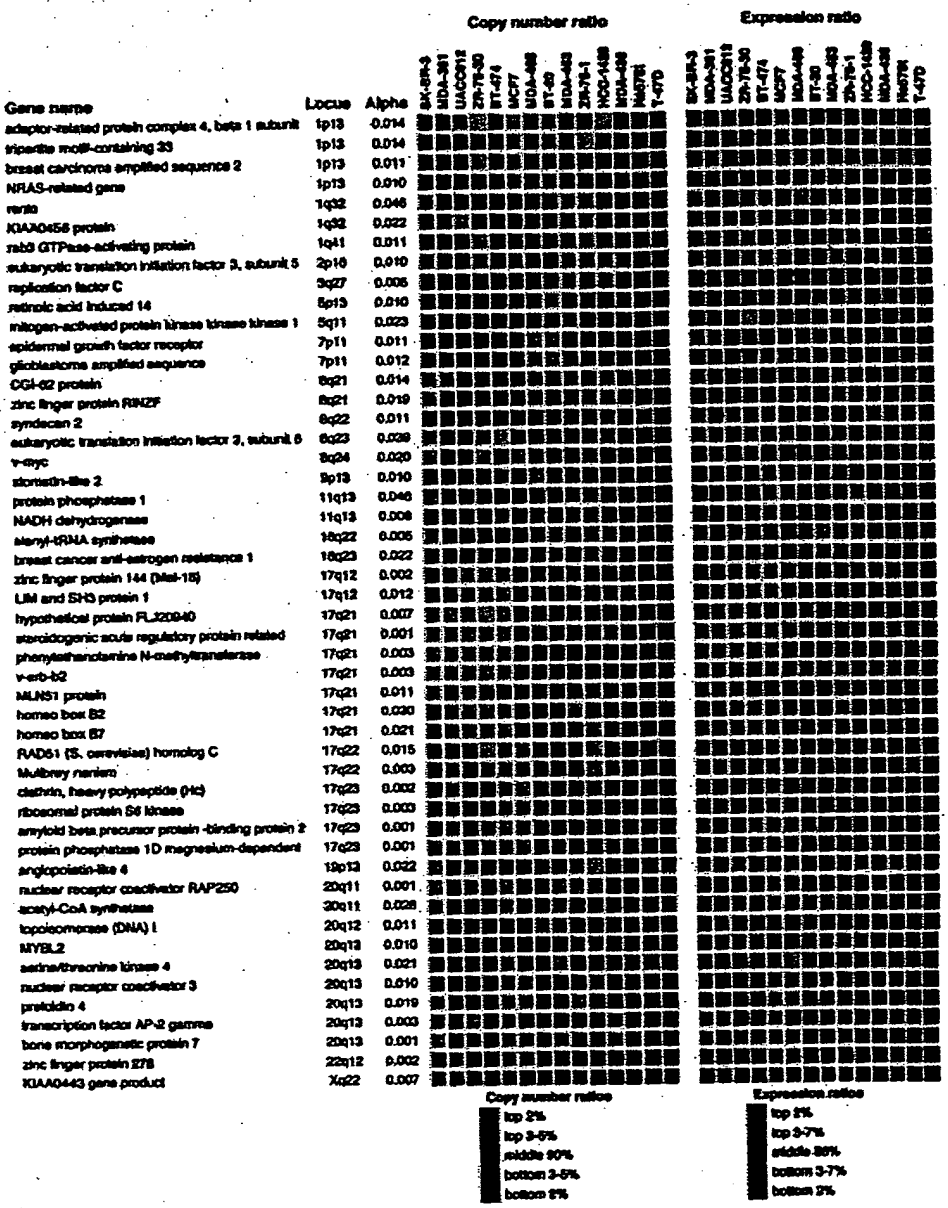


Fig. 4. List of 50 genes with a statistically significant correlation (α value <0.05) between gene copy number and gene expression. Name, chromosomal location, and the α value for each gene are indicated. The genes have been ordered according to their position in the genome. The color maps on the *right* illustrate the copy number and expression ratio patterns in the 14 cell lines. The key to the color code is shown at the *bottom* of the graph. *Gray squares*, missing values. The complete list of 270 genes is shown in supplemental Fig. B.

amplification was validated to be present in 10.2% of 363 primary breast cancers by FISH to a tissue microarray and was associated with poor prognosis of the patients ($P = 0.001$).

Statistical Identification and Characterization of 270 Highly Expressed Genes in Amplicons. Statistical comparison of expression levels of all genes as a function of gene amplification identified 270 genes whose expression was significantly influenced by copy number across all 14 cell lines (Fig. 4, Supplemental Fig. B). According to the gene ontology data,[8] 91 of the 270 genes represented hypothetical proteins or genes with no functional annotation, whereas 179 had associated functional information available. Of these, 151 (84%) are implicated in apoptosis, cell proliferation, signal transduction, and transcription, whereas 28 (16%) had functional annotations that could not be directly linked with cancer.

DISCUSSION

The importance of recurrent gene and chromosome copy number changes in the development and progression of solid tumors has been characterized in >1000 publications applying CGH[9] (9, 10), as well as in a large number of other molecular cytogenetic, cytogenetic, and molecular genetic studies. The effects of these somatic genetic changes on gene expression levels have remained largely unknown, although a few studies have explored gene expression changes occurring in specific amplicons (15, 19–21). Here, we applied genome-wide cDNA microarrays to identify transcripts whose expression changes were attributable to underlying gene copy number alterations in breast cancer.

The overall impact of copy number on gene expression patterns was substantial with the most dramatic effects seen in the case of high-

[8] Internet address: http://www.geneontology.org/.

[9] Internet address: http://www.ncbi.nlm.nih.gov/entrez.

level copy number increase. Low-level copy number gains and losses also had a significant influence on expression levels of genes in the regions affected, but these effects were more subtle on a gene-by-gene basis than those of high-level amplifications. However, the impact of low-level gains on the dysregulation of gene expression patterns in cancer may be equally important if not more important than that of high-level amplifications. Aneuploidy and low-level gains and losses of chromosomal arms represent the most common types of genetic alterations in breast and other cancers and, therefore, have an influence on many genes. Our results in breast cancer extend the recent studies on the impact of aneuploidy on global gene expression patterns in yeast cells, acute myeloid leukemia, and a prostate cancer model system (22–24).

The CGH microarray analysis identified 24 independent breast cancer amplicons. We defined the precise boundaries for many amplicons detected previously by chromosomal CGH (9, 10, 25, 26) and also discovered novel amplicons that had not been detected previously, presumably because of their small size (only 1–2 Mb) or close proximity to other larger amplicons. One of these novel amplicons involved the homeobox gene region at 17q21.3 and led to the overexpression of the *HOXB7* and *HOXB2* genes. The homeodomain transcription factors are known to be key regulators of embryonic development and have been occasionally reported to undergo aberrant expression in cancer (27, 28). *HOXB7* transfection induced cell proliferation in melanoma, breast, and ovarian cancer cells and increased tumorigenicity and angiogenesis in breast cancer (29–32). The present results imply that gene amplification may be a prominent mechanism for overexpressing *HOXB7* in breast cancer and suggest that *HOXB7* contributes to tumor progression and confers an aggressive disease phenotype in breast cancer. This view is supported by our finding of amplification of *HOXB7* in 10% of 363 primary breast cancers, as well as an association of amplification with poor prognosis of the patients.

We carried out a systematic search to identify genes whose expression levels across all 14 cell lines were attributable to amplification status. Statistical analysis revealed 270 such genes (representing ~2% of all genes on the array), including not only previously described amplified genes, such as *HER-2*, *MYC*, *EGFR*, ribosomal protein s6 kinase, and *AIB3*, but also numerous novel genes such as *NRAS-related gene* (1p13), *syndecan-2* (8q22), and *bone morphogenic protein* (20q13.1), whose activation by amplification may similarly promote breast cancer progression. Most of the 270 genes have not been implicated previously in breast cancer development and suggest novel pathogenetic mechanisms. Although we would not expect all of them to be causally involved, it is intriguing that 84% of the genes with associated functional information were implicated in apoptosis, cell proliferation, signal transduction, transcription, or other cellular processes that could directly imply a possible role in cancer progression. Therefore, a detailed characterization of these genes may provide biological insights to breast cancer progression and might lead to the development of novel therapeutic strategies.

In summary, we demonstrate application of cDNA microarrays to the analysis of both copy number and expression levels of over 12,000 transcripts throughout the breast cancer genome, roughly once every 267 kb. This analysis provided: (*a*) evidence of a prominent global influence of copy number changes on gene expression levels; (*b*) a high-resolution map of 24 independent amplicons in breast cancer; and (*c*) identification of a set of 270 genes, the overexpression of which was statistically attributable to gene amplification. Characterization of a novel amplicon at 17q21.3 implicated amplification and overexpression of the *HOXB7* gene in breast cancer, including a clinical association between *HOXB7* amplification and poor patient prognosis. Overall, our results illustrate how the identification of genes activated by gene amplification provides a powerful approach to highlight genes with an important role in cancer as well as to prioritize and validate putative targets for therapy development.

REFERENCES

1. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (Wash. DC), *286:* 531–537, 1999.
2. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature (Lond.), *403:* 503–511, 2000.
3. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature (Lond.), *406:* 536–540, 2000.
4. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.
5. Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. Delineation of prognostic biomarkers in prostate cancer. Nature (Lond.), *412:* 822–826, 2001.
6. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA, *98:* 10869–10874, 2001.
7. Ross, J. S., and Fletcher, J. A. The *HER-2/neu* oncogene: prognostic factor, predictive factor and target for therapy. Semin. Cancer Biol., *9:* 125–138, 1999.
8. Arteaga, C. L. The epidermal growth factor receptor: from mutant oncogene in nonhuman cancers to therapeutic target in human neoplasia. J. Clin. Oncol., *19:* 32–40, 2001.
9. Knuutila, S., Bjorkqvist, A. M., Autio, K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M. L., Tapper, J., Pere, H., El-Rifai, W., *et al.* DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am. J. Pathol., *152:* 1107–1123, 1998.
10. Knuutila S., Autio K., and Aalto Y. Online access to CGH data of DNA sequence copy number changes. Am. J. Pathol., *157:* 689, 2000.
11. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet., *14:* 457–460, 1996.
12. Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res., *6:* 639–645, 1996.
13. Mousses, S., Bittner, M. L., Chen, Y., Dougherty, E. R., Baxevanis, A., Meltzer, P. S., and Trent, J. M. Gene expression analysis by cDNA microarrays. *In:* F. J. Livesey and S. P. Hunt (eds.), Functional Genomics, pp. 113–137. Oxford: Oxford University Press, 2000.
14. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet., *23:* 41–46, 1999.
15. Monni, O., Bärlund, M., Mousses, S., Kononen, J., Sauter, G., Heiskanen, M., Paavola, P., Avela, K., Chen, Y., Bittner, M. L., and Kallioniemi, A. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc. Natl. Acad. Sci. USA, *98:* 5711–5716, 2001.
16. Chen, Y., Dougherty, E. R., and Bittner, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. J. Biomed. Optics, *2:* 364–374, 1997.
17. Bärlund, M., Forozan, F., Kononen, J., Bubendorf, L., Chen, Y., Bittner, M. L., Torhorst, J., Haas, P., Bucher, C., Sauter, G., *et al.* Detecting activation of ribosomal protein S6 kinase by complementary DNA and tissue microarray analysis. J. Natl. Cancer Inst., *92:* 1252–1259, 2000.
18. Andersen, C. L., Hostetter, G., Grigoryan, A., Sauter, G., and Kallioniemi, A. Improved procedure for fluorescence *in situ* hybridization on tissue microarrays. Cytometry, *45:* 83–86, 2001.
19. Kauraniemi, P., Bärlund, M., Monni, O., and Kallioniemi, A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. Cancer Res., *61:* 8235–8240, 2001.
20. Clark, J., Edwards, S., John, M., Flohr, P., Gordon, T., Maillard, K., Giddings, I., Brown, C., Bagherzadeh, A., Campbell, C., Shipley, J., Wooster, R., and Cooper, C. S. Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. Genes Chromosomes Cancer, *34:* 104–114, 2002.
21. Varis, A., Wolf, M., Monni, O., Vakkari, M. L., Kokkola, A., Moskaluk, C., Frierson, H., Powell, S. M., Knuutila, S., Kallioniemi, A., and El-Rifai, W. Targets of gene amplification and overexpression at 17q in gastric cancer. Cancer Res., *62:* 2625–2629, 2002.
22. Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., Friend, S. H., and Marton M. J.

Widespread aneuploidy revealed by DNA microarray expression profiling. Nat. Genet., *25:* 333–337, 2000.

23. Virtaneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de La Chapelle, A., and Krahe, R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc. Natl. Acad. Sci. USA, *98:* 1124–1129, 2001.
24. Phillips, J. L., Hayward, S. W., Wang, Y., Vasselli, J., Pavlovich, C., Padilla-Nash, H., Pezullo, J. R., Ghadimi, B. M., Grossfeld, G. D., Rivera, A., Linehan, W. M., Cunha, G. R., and Ried, T. The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. Cancer Res., *61:* 8143–8149, 2001.
25. Bärlund, M., Tirkkonen, M., Forozan, F., Tanner, M. M., Kallioniemi, O. P., and Kallioniemi, A. Increased copy number at 17q22–q24 by CGH in breast cancer is due to high-level amplification of two separate regions. Genes Chromosomes Cancer, *20:* 372–376, 1997.
26. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Isola, J., Kuukasjärvi, T., Collins, C., Kowbel, D., Guan, X. Y., Trent, J., Gray, J. W., Meltzer, P., and Kallioniemi O. P. Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. Cancer Res., *56:* 3441–3445, 1996.
27. Cillo, C., Faiella, A., Cantile, M., and Boncinelli, E. Homeobox genes and cancer. Exp. Cell Res., *248:* 1–9, 1999.
28. Cillo, C., Cantile, M., Faiella, A., and Boncinelli, E. Homeobox genes in normal and malignant cells. J. Cell. Physiol., *188:* 161–169, 2001.
29. Care, A., Silvani, A., Meccia, E., Mattia, G., Stoppacciaro, A., Parmiani, G., Peschle, C., and Colombo, M. P. HOXB7 constitutively activates basic fibroblast growth factor in melanomas. Mol. Cell. Biol., *16:* 4842–4851, 1996.
30. Care, A., Silvani, A., Meccia, E., Mattia, G., Peschle, C., and Colombo, M. P. Transduction of the SkBr3 breast carcinoma cell line with the *HOXB7* gene induces bFGF expression, increases cell proliferation and reduces growth factor dependence. Oncogene, *16:* 3285–3289, 1998.
31. Care, A., Felicetti, F., Meccia, E., Bottero, L., Parenza, M., Stoppacciaro, A., Peschle, C., and Colombo, M. P. HOXB7: a key factor for tumor-associated angiogenic switch. Cancer Res., *61:* 6532–6539, 2001.
32. Nacra, H., Yang, Y. Q., Montz, F. J., Seidman, J. D., Kurman, R. J., and Roden, R. B. A serologically identified tumor antigen encoded by a homeobox gene promotes growth of ovarian epithelial cells. Proc. Natl. Acad. Sci. USA, *98:* 4060–4065, 2001.

Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors

Jonathan R. Pollack*†‡, Therese Sørlie§, Charles M. Perou¶, Christian A. Rees‖**, Stefanie S. Jeffrey††, Per E. Lonning‡‡, Robert Tibshirani§§, David Botstein‖, Anne-Lise Børresen-Dale§, and Patrick O. Brown†¶¶

Departments of *Pathology, ‖Genetics, ††Surgery, §§Health Research and Policy, and ¶¶Biochemistry, and †Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305; §Department of Genetics, Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway; ‡‡Department of Medicine (Oncology), Haukeland University Hospital, N-5021 Bergen, Norway; and ¶Department of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599

Contributed by Patrick O. Brown, August 6, 2002

Genomic DNA copy number alterations are key genetic events in the development and progression of human cancers. Here we report a genome-wide microarray comparative genomic hybridization (array CGH) analysis of DNA copy number variation in a series of primary human breast tumors. We have profiled DNA copy number alteration across 6,691 mapped human genes, in 44 predominantly advanced, primary breast tumors and 10 breast cancer cell lines. While the overall patterns of DNA amplification and deletion corroborate previous cytogenetic studies, the high-resolution (gene-by-gene) mapping of amplicon boundaries and the quantitative analysis of amplicon shape provide significant improvement in the localization of candidate oncogenes. Parallel microarray measurements of mRNA levels reveal the remarkable degree to which variation in gene copy number contributes to variation in gene expression in tumor cells. Specifically, we find that 62% of highly amplified genes show moderately or highly elevated expression, that DNA copy number influences gene expression across a wide range of DNA copy number alterations (deletion, low-, mid- and high-level amplification), that on average, a 2-fold change in DNA copy number is associated with a corresponding 1.5-fold change in mRNA levels, and that overall, at least 12% of all the variation in gene expression among the breast tumors is directly attributable to underlying variation in gene copy number. These findings provide evidence that widespread DNA copy number alteration can lead directly to global deregulation of gene expression, which may contribute to the development or progression of cancer.

Conventional cytogenetic techniques, including comparative genomic hybridization (CGH) (1), have led to the identification of a number of recurrent regions of DNA copy number alteration in breast cancer cell lines and tumors (2–4). While some of these regions contain known or candidate oncogenes [e.g., FGFR1 (8p11), MYC (8q24), CCND1 (11q13), ERBB2 (17q12), and ZNF217 (20q13)] and tumor suppressor genes [RB1 (13q14) and TP53 (17p13)], the relevant gene(s) within other regions (e.g., gain of 1q, 8q22, and 17q22–24, and loss of 8p) remain to be identified. A high-resolution genome-wide map, delineating the boundaries of DNA copy number alterations in tumors, should facilitate the localization and identification of oncogenes and tumor suppressor genes in breast cancer. In this study, we have created such a map, using array-based CGH (5–7) to profile DNA copy number alteration in a series of breast cancer cell lines and primary tumors.

An unresolved question is the extent to which the widespread DNA copy number changes that we and others have identified in breast tumors alter expression of genes within involved regions. Because we had measured mRNA levels in parallel in the same samples (8), using the same DNA microarrays, we had an opportunity to explore on a genomic scale the relationship between DNA copy number changes and gene expression. From this analysis, we have identified a significant impact of widespread DNA copy number alteration on the transcriptional programs of breast tumors.

Materials and Methods

Tumors and Cell Lines. Primary breast tumors were predominantly large (>3 cm), intermediate-grade, infiltrating ductal carcinomas, with more than 50% being lymph node positive. The fraction of tumor cells within specimens averaged at least 50%. Details of individual tumors have been published (8, 9), and are summarized in Table 1, which is published as supporting information on the PNAS web site, www.pnas.org. Breast cancer cell lines were obtained from the American Type Culture Collection. Genomic DNA was isolated either using Qiagen genomic DNA columns, or by phenol/chloroform extraction followed by ethanol precipitation.

DNA Labeling and Microarray Hybridizations. Genomic DNA labeling and hybridizations were performed essentially as described in Pollack *et al.* (7), with slight modifications. Two micrograms of DNA was labeled in a total volume of 50 microliters and the volumes of all reagents were adjusted accordingly. "Test" DNA (from tumors and cell lines) was fluorescently labeled (Cy5) and hybridized to a human cDNA microarray containing 6,691 different mapped human genes (i.e., UniGene clusters). The "reference" (labeled with Cy3) for each hybridization was normal female leukocyte DNA from a single donor. The fabrication of cDNA microarrays and the labeling and hybridization of mRNA samples have been described (8).

Data Analysis and Map Positions. Hybridized arrays were scanned on a GenePix scanner (Axon Instruments, Foster City, CA), and fluorescence ratios (test/reference) calculated using SCANALYZE software (available at http://rana.lbl.gov). Fluorescence ratios were normalized for each array by setting the average log fluorescence ratio for all array elements equal to 0. Measurements with fluorescence intensities more than 20% above background were considered reliable. DNA copy number profiles that deviated significantly from background ratios measured in normal genomic DNA control hybridizations were interpreted as evidence of real DNA copy number alteration (see *Estimating Significance of Altered Fluorescence Ratios* in the supporting information). When indicated, DNA copy number profiles are displayed as a moving average (symmetric 5-nearest neighbors). Map positions for arrayed human cDNAs were assigned by

Abbreviation: CGH, comparative genomic hybridization.

†To whom reprint requests should be addressed at: Department of Pathology, Stanford University School of Medicine, CCSR Building, Room 3245A, 269 Campus Drive, Stanford, CA 94305-5176. E-mail: pollack1@stanford.edu.

**Present address: Zyomyx Inc., Hayward, CA 94545.

GENETICS

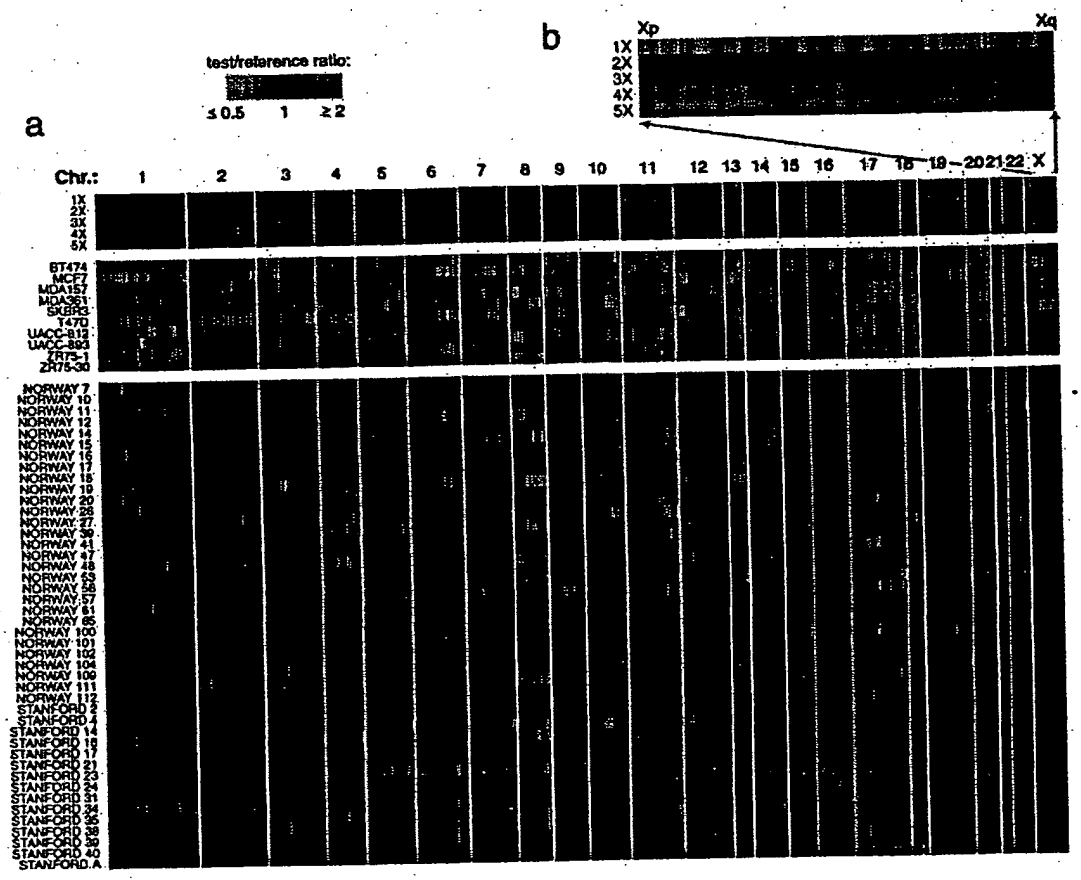
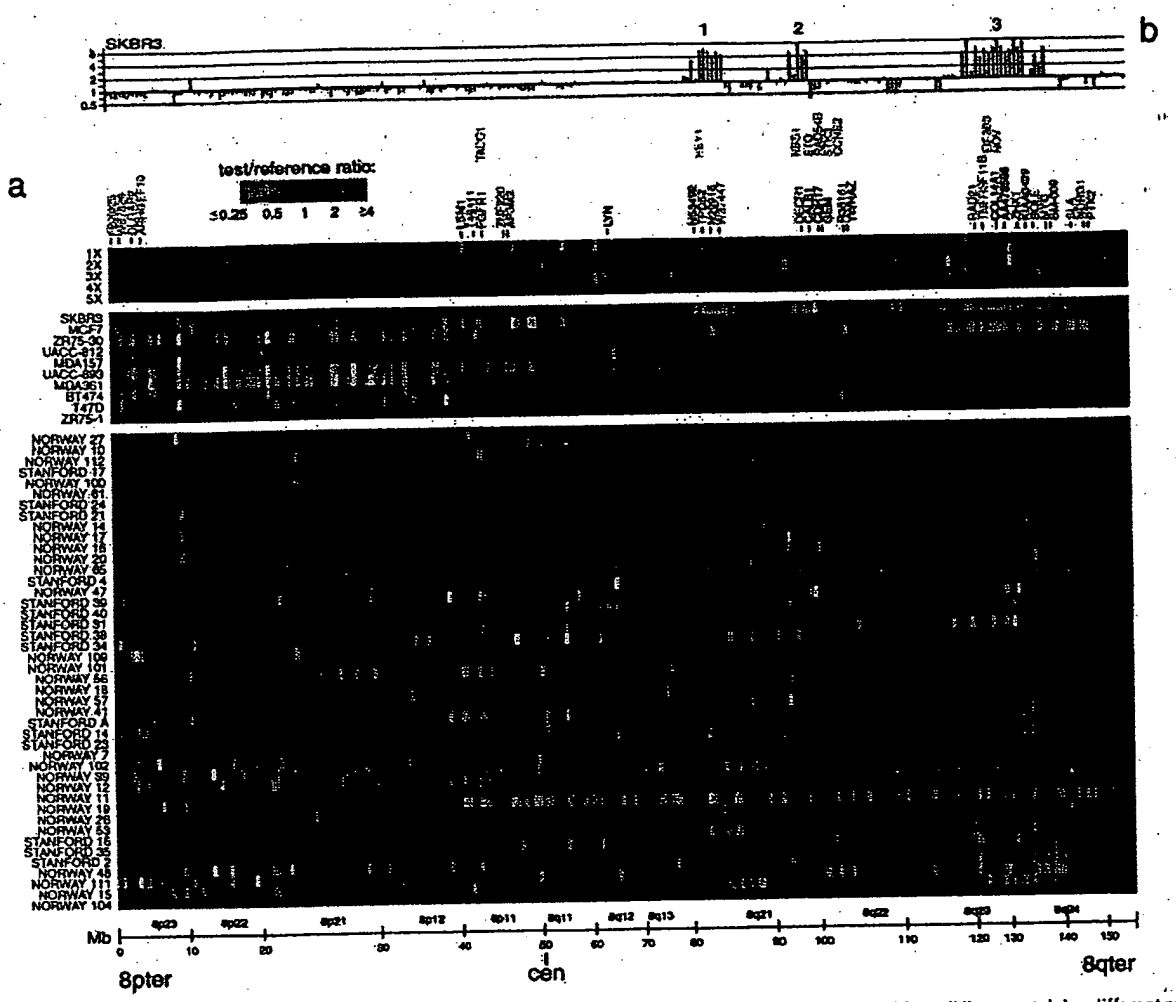


Fig. 1. Genome-wide measurement of DNA copy number alteration by array CGH. (a) DNA copy number profiles are illustrated for cell lines containing different numbers of X chromosomes, for breast cancer cell lines, and for breast tumors. Each row represents a different cell line or tumor, and each column represents one of 6,691 different mapped human genes present on the microarray, ordered by genome map position from 1pter through Xqter. Moving average (symmetric 5-nearest neighbors) fluorescence ratios (test/reference) are depicted using a $\log_2$-based pseudocolor scale (indicated), such that red luminescence reflects fold-amplification, green luminescence reflects fold-deletion, and black indicates no change (gray indicates poorly measured data). (b) Enlarged view of DNA copy number profiles across the X chromosome, shown for cell lines containing different numbers of X chromosomes.

identifying the starting position of the best and longest match of any DNA sequence represented in the corresponding UniGene cluster (10) against the "Golden Path" genome assembly (http://genome.ucsc.edu/; Oct 7, 2000 Freeze). For UniGene clusters represented by multiple arrayed elements, mean fluorescence ratios (for all elements representing the same UniGene cluster) are reported. For mRNA measurements, fluorescence ratios are "mean-centered" (i.e., reported relative to the mean ratio across the 44 tumor samples). The data set described here can be accessed in its entirety in the supporting information.

Results

We performed CGH on 44 predominantly locally advanced, primary breast tumors and 10 breast cancer cell lines, using cDNA microarrays containing 6,691 different mapped human genes (Fig. 1*a*; also see *Materials and Methods* for details of microarray hybridizations). To take full advantage of the improved spatial resolution of array CGH, we ordered (fluorescence ratios for) the 6,691 cDNAs according to the "Golden Path" (http://genome.ucsc.edu/) genome assembly of the draft human genome sequences (11). In so doing, arrayed cDNAs not only themselves represent genes of potential interest (e.g., candidate oncogenes within amplicons), but also provide precise genetic landmarks for chromosomal regions of amplification and deletion. Parallel analysis of DNA from cell lines containing different numbers of X chromosomes (Fig. 1*b*), as we did before (7), demonstrated the sensitivity of our method to detect single-copy loss (45, XO), and 1.5- (47,XXX), 2- (48,XXXX), or 2.5-fold (49,XXXXX) gains (also see Fig. 5, which is published as supporting information on the PNAS web site). Fluorescence ratios were linearly proportional to copy number ratios, which were slightly underestimated, in agreement with previous observations (7). Numerous DNA copy number alterations were evident in both the breast cancer cell lines and primary tumors (Fig. 1*a*), detected in the tumors despite the presence of euploid non-tumor cell types; the magnitudes of the observed changes were generally lower in the tumor samples. DNA copy-number alterations were found in every cancer cell line and tumor, and on every human chromosome in at least one sample. Recurrent regions of DNA copy number gain and loss were readily identifiable. For example, gains within 1q, 8q, 17q, and 20q were observed in a high proportion of breast cancer cell lines/tumors (90%/69%, 100%/47%, 100%/60%, and 90%/44%, respectively), as were losses within 1p, 3p, 8p, and 13q (80%/24%, 80%/22%, 80%/22%, and 70%/18%, respectively), consistent with published cytogenetic studies (refs. 2–4; a complete listing of gains/losses is provided in Tables 2 and 3, which are published as supporting information on the PNAS web site). The total



BEST AVAILABLE COPY

Fig. 2. DNA copy number alteration across chromosome 8 by array CGH. (a) DNA copy number profiles are illustrated for cell lines containing different numbers of X chromosomes, for breast cancer cell lines, and for breast tumors. Breast cancer cell lines and tumors are separately ordered by hierarchical clustering to highlight recurrent copy number changes. The 241 genes present on the microarrays and mapping to chromosome 8 are ordered by position along the chromosome. Fluorescence ratios (test/reference) are depicted by a $\log_2$ pseudocolor scale (indicated). Selected genes are indicated with color-coded text (red, increased; green, decreased; black, no change; gray, not well measured) to reflect correspondingly altered mRNA levels (observed in the majority of the subset of samples displaying the DNA copy number change). The map positions for genes of interest that are not represented on the microarray are indicated in the row above those genes represented on the array. (b) Graphical display of DNA copy number profile for breast cancer cell line SKBR3. Fluorescence ratios (tumor/normal) are plotted on a $\log_2$ scale for chromosome 8 genes, ordered along the chromosome.

number of genomic alterations (gains and losses) was found to be significantly higher in breast tumors that were high grade (P = 0.008), consistent with published CGH data (3), estrogen receptor negative (P = 0.04), and harboring TP53 mutations (P = 0.0006) (see Table 4, which is published as supporting information on the PNAS web site).

The improved spatial resolution of our array CGH analysis is illustrated for chromosome 8, which displayed extensive DNA copy number alteration in our series. A detailed view of the variation in the copy number of 241 genes mapping to chromosome 8 revealed multiple regions of recurrent amplification; each of these potentially harbors a different known or previously uncharacterized oncogene (Fig. 2*a*). The complexity of amplicon structure is most easily appreciated in the breast cancer cell line SKBR3. Although a conventional CGH analysis of 8q in SKBR3 identified only two distinct regions of amplification (12), we observed three distinct regions of high-level amplification (labeled 1–3 in Fig. 2*b*). For each of these regions we can define the boundaries of the interval recurrently amplified in the tumors we examined; in each case, known or plausible candidate oncogenes can be identified (a description of these regions, as well as the recurrently amplified regions on chromosomes 17 and 20, can be found in Figs. 6 and 7, which are published as supporting information on the PNAS web site).

For a subset of breast cancer cell lines and tumors (4 and 37, respectively), and a subset of arrayed genes (6,095), mRNA levels were quantitatively measured in parallel by using cDNA microarrays (8). The parallel assessment of mRNA levels is useful in the interpretation of DNA copy number changes. For example, the highly amplified genes that are also highly expressed are the strongest candidate oncogenes within an amplicon. Perhaps more significantly, our parallel analysis of DNA copy number changes and mRNA levels provides us the opportunity to assess the global impact of widespread DNA copy number alteration on gene expression in tumor cells.

A strong influence of DNA copy number on gene expression is evident in an examination of the pseudocolor representations

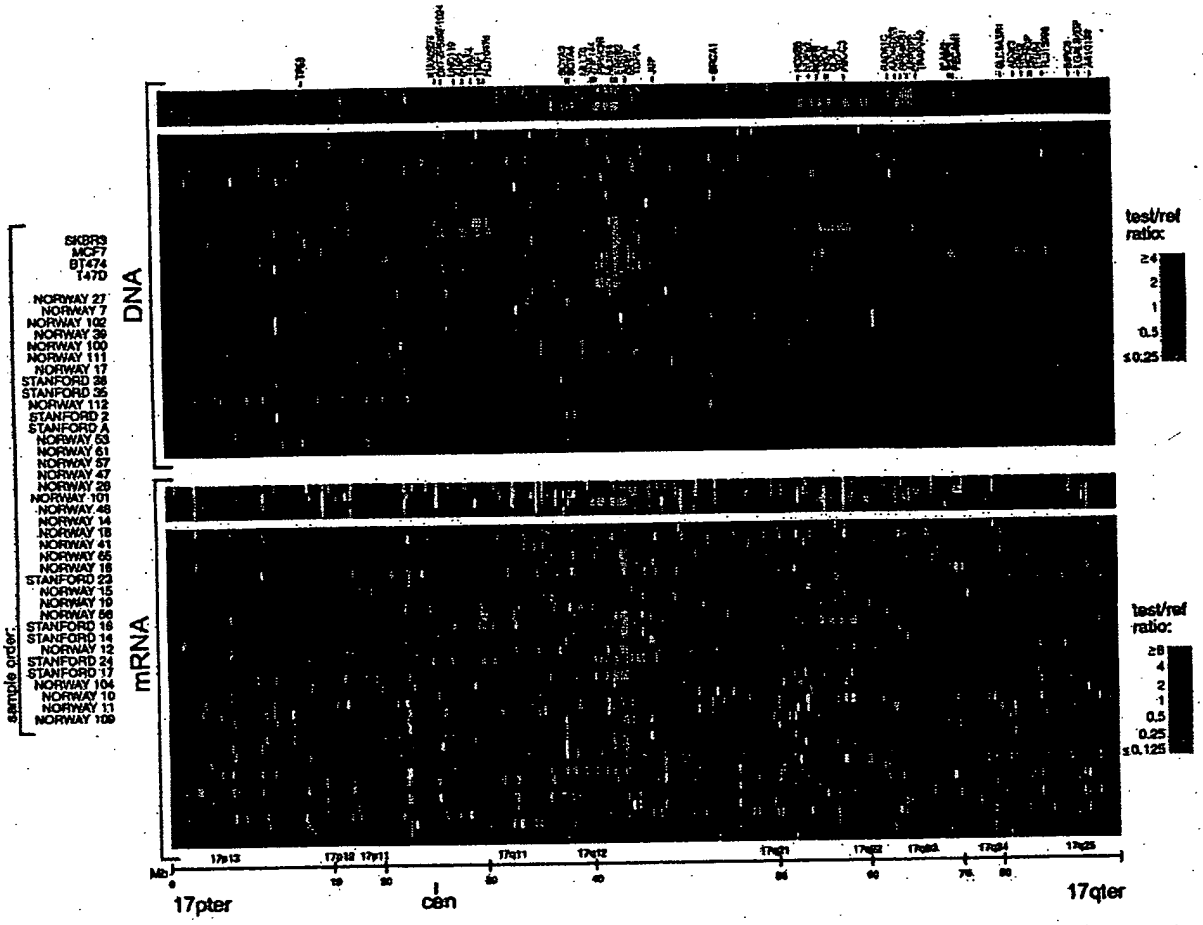


Fig. 3. Concordance between DNA copy number and gene expression across chromosome 17. DNA copy number alteration (*Upper*) and mRNA levels (*Lower*) are illustrated for breast cancer cell lines and tumors. Breast cancer cell lines and tumors are separately ordered by hierarchical clustering (*Upper*), and the identical sample order is maintained (*Lower*). The 354 genes present on the microarrays and mapping to chromosome 17, and for which both DNA copy number and mRNA levels were determined, are ordered by position along the chromosome; selected genes are indicated in color-coded text (see Fig. 2 legend). Fluorescence ratios (test/reference) are depicted by separate $\log_2$ pseudocolor scales (indicated).

of DNA copy number and mRNA levels for genes on chromosome 17 (Fig. 3). The overall patterns of gene amplification and elevated gene expression are quite concordant; i.e., a significant fraction of highly amplified genes appear to be correspondingly highly expressed. The concordance between high-level amplification and increased gene expression is not restricted to chromosome 17. Genome-wide, of 117 high-level DNA amplifications (fluorescence ratios >4, and representing 91 different genes), 62% (representing 54 different genes; see Table 5, which is published as supporting information on the PNAS web site) are found associated with at least moderately elevated mRNA levels (mean-centered fluorescence ratios >2), and 42% (representing 36 different genes) are found associated with comparably highly elevated mRNA levels (mean-centered fluorescence ratios >4).

To determine the extent to which DNA deletion and lower-level amplification (in addition to high-level amplification) are also associated with corresponding alterations in mRNA levels, we performed three separate analyses on the complete data set (4 cell lines and 37 tumors, across 6,095 genes). First, we determined the average mRNA levels for each of five classes of genes, representing DNA deletion, no change, and low-, medium-, and high-level amplification (Fig. 4*a*). For both the breast cancer cell lines and tumors, average mRNA levels tracked with DNA copy number across all five classes, in a statistically significant fashion (P values for pair-wise Student's t tests comparing adjacent classes: cell lines, 4×10^{-49}, 1×10^{-49}, 5×10^{-5}, 1×10^{-2}; tumors, 1×10^{-43}, 1×10^{-214}, 5×10^{-41}, 1×10^{-4}). A linear regression of the average log(DNA copy number), for each class, against average log(mRNA level) demonstrated that on average, a 2-fold change in DNA copy number was accompanied by 1.4- and 1.5-fold changes in mRNA level for the breast cancer cell lines and tumors, respectively (Fig. 4*a*, regression line not shown). Second, we characterized the distribution of the 6,095 correlations between DNA copy number and mRNA level, each across the 37 tumor samples (Fig. 4*b*). The distribution of correlations forms a normal-shaped curve, but with the peak markedly shifted in the positive direction from zero. This shift is statistically significant, as evidenced in a plot of observed vs. expected correlations (Fig. 4*c*), and reflects a pervasive global influence of DNA copy number alterations on gene expression. Notably, the highest correlations between DNA copy number and mRNA level (the right tail of the distribution in Fig. 4*b*) comprise both amplified and deleted genes (data not shown). Third, we used a linear regression model to estimate the fraction of all variation measured in mRNA levels among the 37

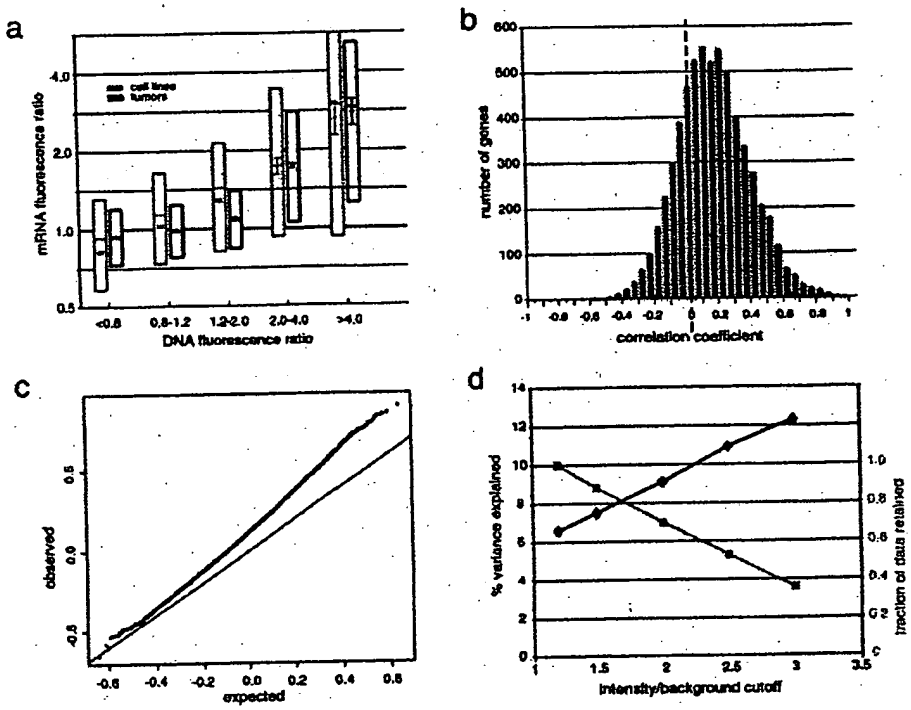


Fig. 4. Genome-wide influence of DNA copy number alterations on mRNA levels. (*a*) For breast cancer cell lines (gray) and tumor samples (black), both mean-centered mRNA fluorescence ratio ($\log_2$ scale) quartiles (box plots indicate 25th, 50th, and 75th percentile) and averages (diamonds; *Y*-value error bars indicate standard errors of the mean) are plotted for each of five classes of genes, representing DNA deletion (tumor/normal ratio < 0.8), no change (0.8–1.2), low- (1.2–2), medium- (2–4), and high-level (>4) amplification. *P* values for pair-wise Student's t tests, comparing averages between adjacent classes (moving left to right), are 4×10^{-49}, 1×10^{-49}, 5×10^{-5}, 1×10^{-2} (cell lines); and 1×10^{-43}, 1×10^{-214}, 5×10^{-41}, 1×10^{-4} (tumors). (*b*) Distribution of correlations between DNA copy number and mRNA levels, for 6,095 different human genes across 37 breast tumor samples. (*c*) Plot of observed versus expected correlation coefficients. The expected values were obtained by randomization of the sample labels in the DNA copy number data set. The line of unity is indicated. (*d*) Percent variance in gene expression (among tumors) directly explained by variation in gene copy number. Percent variance explained (black line) and fraction of data retained (gray line) are plotted for different fluorescence intensity/background (a rough surrogate for signal/noise) cutoff values. Fraction of data retained is relative to the 1.2 intensity/background cutoff. Details of the linear regression model used to estimate the fraction of variation in gene expression attributable to underlying DNA copy number alteration can be found in the supporting information (see *Estimating the Fraction of Variation in Gene Expression Attributable to Underlying DNA Copy Number Alteration*).

tumors that could be attributed to underlying variation in DNA copy number. From this analysis, we estimate that, overall, about 7% of all of the observed variation in mRNA levels can be explained directly by variation in copy number of the altered genes (Fig. 4*d*). We can reduce the effects of experimental measurement error on this estimate by using only that fraction of the data most reliably measured (fluorescence intensity/background >3); using that data, our estimate of the percent variation in mRNA levels directly attributed to variation in gene copy number increases to 12% (Fig. 4*d*). This still undoubtedly represents a significant underestimate, as the observed variation in global gene expression is affected not only by true variation in the expression programs of the tumor cells themselves, but also by the variable presence of non-tumor cell types within clinical samples.

Discussion

This genome-wide, array CGH analysis of DNA copy number alteration in a series of human breast tumors demonstrates the usefulness of defining amplicon boundaries at high resolution (gene-by-gene), and quantitatively measuring amplicon shape, to assist in locating and identifying candidate oncogenes. By analyzing mRNA levels in parallel, we have also discovered that changes in DNA copy number have a large, pervasive, direct effect on global gene expression patterns in both breast cancer cell lines and tumors. Although the DNA microarrays used in our analysis may display a bias toward characterized and/or highly expressed genes, because we are examining such a large fraction of the genome (approximately 20% of all human genes), and because, as detailed above, we are likely underestimating the contribution of DNA copy number changes to altered gene expression, we believe our findings are likely to be generalizable (but would nevertheless still be remarkable if only applicable to this set of ~6,100 genes).

In budding yeast, aneuploidy has been shown to result in chromosome-wide gene expression biases (13). Two recent studies have begun to examine the global relationship between DNA copy number and gene expression in cancer cells. In agreement with our findings, Phillips *et al.* (14) have shown that with the acquisition of tumorigenicity in an immortalized prostate epithelial cell line, new chromosomal gains and losses resulted in a statistically significant respective increase and decrease in the average expression level of involved genes. In contrast, Platzer *et al.* (15) recently reported that in metastatic colon tumors only ~4% of genes within amplified regions were found more highly (>2-fold) expressed, when compared with normal colonic epithelium. This report differs substantially from our finding that 62% of highly amplified genes in breast cancer exhibit at least 2-fold increased expression. These contrasting findings may reflect methodological differences between the

GENETICS

studies. For example, the study of Platzer *et al.* (15) may have systematically under-measured gene expression changes. In this regard it is remarkable that only 14 transcripts of many thousand residing within unamplified chromosomal regions were found to exhibit at least 4-fold altered expression in metastatic colon cancer. Additionally, their reliance on lower-resolution chromosomal CGH may have resulted in poorly delimiting the boundaries of high-complexity amplicons, effectively overcalling regions with amplification. Alternatively, the contrasting findings for amplified genes may represent real biological differences between breast and metastatic colon tumors; resolution of this issue will require further studies.

Our finding that widespread DNA copy number alteration has a large, pervasive and direct effect on global gene expression patterns in breast cancer has several important implications. First, this finding supports a high degree of copy number-dependent gene expression in tumors. Second, it suggests that most genes are not subject to specific autoregulation or dosage compensation. Third, this finding cautions that elevated expression of an amplified gene cannot alone be considered strong independent evidence of a candidate oncogene's role in tumorigenesis. In our study, fully 62% of highly amplified genes demonstrated moderately or highly elevated expression. This highlights the importance of high-resolution mapping of amplicon boundaries and shape [to identify the "driving" gene(s) within amplicons (16)], on a large number of samples, in addition to functional studies. Fourth, this finding suggests that analyzing the genomic distribution of expressed genes, even within existing microarray gene expression data sets, may permit the inference of DNA copy number aberration, particularly aneuploidy (where gene expression can be averaged across large chromosomal regions; see Fig. 3 and supporting information). Fifth, this finding implies that a substantial portion of the phenotypic uniqueness (and by extension, the heterogeneity in clinical behavior) among patients' tumors may be traceable to underlying variation in DNA copy number. Sixth, this finding supports a possible role for widespread DNA copy number alteration in tumorigenesis (17, 18), beyond the amplification of specific oncogenes and deletion of specific tumor suppressor genes. Widespread DNA copy number alteration, and the concomitant widespread imbalance in gene expression, might disrupt critical stochiometric relationships in cell metabolism and physiology (e.g., proteosome, mitotic spindle), possibly promoting further chromosomal instability and directly contributing to tumor development or progression. Finally, our findings suggest the possibility of cancer therapies that exploit specific or global imbalances in gene expression in cancer.

We thank the many members of the P.O.B. and D.B. labs for helpful discussions. J.R.P. was a Howard Hughes Medical Institute Physician Postdoctoral Fellow during a portion of this work. P.O.B. is a Howard Hughes Medical Institute Associate Investigator. This work was supported by grants from the National Institutes of Health, the Howard Hughes Medical Institute, the Norwegian Cancer Society, and the Norwegian Research Council.

1. Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F. & Pinkel, D. (1992) *Science* 258, 818–821.
2. Kallioniemi, A., Kallioniemi, O. P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H. S., Pinkel, D., Gray, J. W. & Waldman, F. M. (1994) *Proc. Natl. Acad. Sci. USA* 91, 2156–2160.
3. Tirkkonen, M., Tanner, M., Karhu, R., Kallioniemi, A., Isola, J. & Kallioniemi, O. P. (1998) *Genes Chromosomes Cancer* 21, 177–184.
4. Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., Gooden, G. C., Ethier, S. P., Kallioniemi, A. & Kallioniemi, O. P. (2000) *Cancer Res.* 60, 4519–4525.
5. Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Lichter, P. (1997) *Genes Chromosomes Cancer* 20, 399–407.
6. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., *et al.* (1998) *Nat. Genet.* 20, 207–211.
7. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. & Brown, P. O. (1999) *Nat. Genet.* 23, 41–46.
8. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* (2000) *Nature (London)* 406, 747–752.
9. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* 98, 10869–10874.
10. Schuler, G. D. (1997) *J. Mol. Med.* 75, 694–698.
11. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature (London)* 409, 860–921.
12. Fejzo, M. S., Godfrey, T., Chen, C., Waldman, F. & Gray, J. W. (1998) *Genes Chromosomes Cancer* 22, 105–113.
13. Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., *et al.* (2000) *Nat. Genet.* 25, 333–337.
14. Phillips, J. L., Hayward, S. W., Wang, Y., Vasselli, J., Pavlovich, C., Padilla-Nash, H., Pezullo, J. R., Ghadimi, B. M., Grossfeld, G. D., Rivera, A., *et al.* (2001) *Cancer Res.* 61, 8143–8149.
15. Platzer, P., Upender, M. B., Wilson, K., Willis, J., Lutterbaugh, J., Nosrati, A., Willson, J. K., Mack, D., Ried, T. & Markowitz, S. (2002) *Cancer Res.* 62, 1134–1138.
16. Albertson, D. G., Ylstra, B., Segraves, R., Collins, C., Dairkee, S. H., Kowbel, D., Kuo, W. L., Gray, J. W. & Pinkel, D. (2000) *Nat. Genet.* 25, 144–146.
17. Li, R., Yerganian, G., Duesberg, P., Kraemer, A., Willer, A., Rausch, C. & Hehlmann, R. (1997) *Proc. Natl. Acad. Sci. USA* 94, 14506–14511.
18. Rasnick, D. & Duesberg, P. H. (1999) *Biochem. J.* 340, 621–630.

TECHNICAL UPDATE

FROM YOUR LABORATORY SERVICES PROVIDER

HER-2/neu Breast Cancer Predictive Testing

Julie Sanford Hanna, Ph.D. and Dan Mornin, M.D.

EACH YEAR, OVER 182,000 WOMEN in the United States are diagnosed with breast cancer, and approximately 45,000 die of the disease.[1] Incidence appears to be increasing in the United States at a rate of roughly 2% per year. The reasons for the increase are unclear, but non-genetic risk factors appear to play a large role.[2]

Five-year survival rates range from approximately 65%-85%, depending on demographic group, with a significant percentage of women experiencing recurrence of their cancer within 10 years of diagnosis. One of the factors most predictive for recurrence once a diagnosis of breast cancer has been made is the number of axillary lymph nodes to which tumor has metastasized. Most node-positive women are given adjuvant therapy, which increases their survival. However, 20%-30% of patients without axillary node involvement also develop recurrent disease, and the difficulty lies in how to identify this high-risk subset of patients. These patients could benefit from increased surveillance, early intervention, and treatment.

Prognostic markers currently used in breast cancer recurrence prediction include tumor size, histological grade, steroid hormone receptor status, DNA ploidy, proliferative index, and cathepsin D status. Expression of growth factor receptors and over-expression of the HER-2/neu oncogene have also been identified as having value regarding treatment regimen and prognosis.

HER-2/neu (also known as c-erbB2) is an oncogene that encodes a transmembrane glycoprotein that is homologous to, but distinct from, the epidermal growth factor receptor. Numerous studies have indicated that high levels of expression of this protein are associated with rapid tumor growth, certain forms of therapy resistance, and shorter disease-free survival. The gene has been shown to be amplified and/or overexpressed in 10%-30% of invasive breast cancers and in 40%-60% of intraductal breast carcinoma.[3]

There are two distinct FDA-approved methods by which HER-2/neu status can be evaluated: immunohistochemistry (IHC, HercepTest™) and FISH (fluorescent in situ hybridization, PathVysion™ Kit). Both methods can be performed on archived and current specimens. The first method allows visual assessment of the amount of HER-2/neu protein present on the cell membrane. The latter method allows direct quantification of the level of gene amplification present in the tumor, enabling differentiation between low- versus high-amplification. At least one study has demonstrated a difference in recurrence risk in women younger than 40 years of age for low- versus high-amplified tumors (54.5% compared to 85.7%); this is compared to a recurrence rate of 16.7% for patients with no HER-2/neu gene amplification.[4] HER-2/neu status may be particularly important to establish in women with small ($\leq$1 cm) tumor size.

The choice of methodology for determination of HER-2/neu status depends in part on the clinical setting. FDA approval for the Vysis FISH test was granted based on clinical trials involving 1549 node-positive patients. Patients received one of three different treatments consisting of different doses of cyclophosphamide, Adriamycin, and 5-fluorouracil (CAF). The study showed that patients with amplified HER-2/neu benefited from treatment with higher doses of adriamycin-based therapy, while those with normal HER-2/neu levels did not. The study therefore identified a sub-set of women, who because they did not benefit from more aggressive treatment, did not need to be exposed to the associated side effects. In addition, other evidence indicates that HER-2/neu amplification in node-negative patients can be used as an independent prognostic indicator for early recurrence, recurrent disease at any time and disease-related death.[5] Demonstration of HER-2/neu gene amplification by FISH has also been shown to be of value in predicting response to chemotherapy in stage-2 breast cancer patients.

Selection of patients for Herceptin© (Trastuzumab) monoclonal antibody therapy, however, is based upon demonstration of HER-2/neu protein overexpression using HercepTest™. Studies using Herceptin© in patients with metastatic breast cancer show an increase in time to disease progression, increased response rate to chemotherapeutic agents and a small increase in overall survival rate. The FISH assays have not yet been approved for this purpose, and studies looking at response to Herceptin© in patients with or without gene amplification status determined by FISH are in progress.

In general, FISH and IHC results correlate well. However, subsets of tumors are found which show discordant results; i.e., protein overexpression without gene amplification or lack of protein overexpression with gene amplification. The clinical significance of such results is unclear. Based on the above considerations, HER-2/neu testing at SHMC/PAML will utilize immunohistochemistry (HercepTest©) as a screen, followed by FISH in IHC-negative cases. Alternatively, either method may be ordered individually depending on the clinical setting or clinician preference.

CPT code information

HER-2/neu via IHC

88342 (including interpretive report)

HER-2/neu via FISH

88271×2 Molecular cytogenetics, DNA probe, each

88274 Molecular cytogenetics, interphase in situ hybridization, analyze 25-99 cells

88291 Cytogenetics and molecular cytogenetics, interpretation and report

Procedural Information

Immunohistochemistry is performed using the FDA-approved DAKO antibody kit, Herceptest©. The DAKO kit contains reagents required to complete a two-step immunohistochemical staining procedure for routinely processed, paraffin-embedded specimens. Following incubation with the primary rabbit antibody to human HER-2/neu protein, the kit employs a ready-to-use dextran-based visualization reagent. This reagent consists of both secondary goat anti-rabbit antibody molecules with horseradish peroxidase molecules linked to a common dextran polymer backbone, thus eliminating the need for sequential application of link antibody and peroxidase conjugated antibody. Enzymatic conversion of the subsequently added chromogen results in formation of visible reaction product at the antigen site. The specimen is then counterstained; a pathologist using light-microscopy interprets results.

FISH analysis at SHMC/PAML is performed using the FDA-approved PathVysion™ HER-2/neu DNA probe kit, produced by Vysis, Inc. Formalin fixed, paraffin-embedded breast tissue is processed using routine histological methods, and then slides are treated to allow hybridization of DNA probes to the nuclei present in the tissue section. The Pathvysion™ kit contains two direct-labeled DNA probes, one specific for the alphoid repetitive DNA (CEP 17, spectrum orange) present at the chromosome 17 centromere and the second for the HER-2/neu oncogene located at 17q11.2-12 (spectrum green). Enumeration of the probes allows a ratio of the number of copies of chromosome 17 to the number of copies of HER-2/neu to be obtained; this enables quantification of low versus high amplification levels, and allows an estimate of the percentage of cells with HER-2/neu gene amplification. The clinically relevant distinction is whether the gene amplification is due to increased gene copy number on the two chromosome 17 homologues normally present or an increase in the number of chromosome 17s in the cells. In the majority of cases, ratio equivalents less than 2.0 are indicative of a normal/negative result, ratios of 2.1 and over indicate that amplification is present and to what degree. Interpretation of this data will be performed and reported from the Vysis-certified Cytogenetics laboratory at SHMC.

References

1. Wingo, P.A., Tong, T., Bolden, S., "Cancer Statistics", 1995;45:1:8-31.
2. "Cancer Rates and Risks", 4th ed., National Institutes of Health, National Cancer Institute, 1996, p. 120.
3. Slamon, D.J., Clark, G.M., Song, S.G., Levin, W.J., Ullrich, A., McGuire, W.L. "Human breast Cancer: Correlation of relapse and survival with amplification of the her-2/neu oncogene". Science, 235:177-182, 1987.
4. Xing, W.R., Gilchrist, K.W., Harris, C.P., Samson, W., Meisner, L.F. "FISH detection of HER-s/neu oncogene amplification in early onset breast cancer". Breast Cancer Res. And Treatment 39(2):203-212, 1996.
5. Press, M.F. Bernstein, L., Thomas, P.A., Meisner, L.F., Zhou, J.Y., Ma, Y., Hung, G., Robinson, R.A., Harris, C., El-Naggar, A., Slamon, D.J., Phillips, R.N., Ross, J.S., Wolman, S.R., Flom, K.J., "Her-2/neu gene amplification characterized by fluorescence in situ hybridization: poor prognosis in node-negative breast carcinomas", J. Clinical Oncology 15(8):2894-2904, 1997.

Provided for the clients of

PATHOLOGY ASSOCIATES MEDICAL LABORATORIES
PACLAB NETWORK LABORATORIES
TRI-CITIES LABORATORY
TREASURE VALLEY LABORATORY

For more information, please contact your local representative.

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Ashkenazi et al. Serial No.: 09/903,925 Filed: July 11, 2001 For: SECRETED AND TRANSMEMBRANE POLYPEPTIDES AND NUCLEIC ACIDS	Group Art Unit: 1647 Examiner: Fozia Hamid CERTIFICATE OF MAILING I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Assistant Commissioner of Patents, Washington, D.C. 20231 on Date

DECLARATION OF AUDREY D. GODDARD, Ph.D UNDER 37 C.F.R. § 1.132

Assistant Commissioner of Patents
Washington, D.C. 20231

Sir:

I, Audrey D. Goddard, Ph.D. do hereby declare and say as follows:

1. I am a Senior Clinical Scientist at the Experimental Medicine/BioOncology, Medical Affairs Department of Genentech, Inc., South San Francisco, California 94080.

2. Between 1993 and 2001, I headed the DNA Sequencing Laboratory at the Molecular Biology Department of Genentech, Inc. During this time, my responsibilities included the identification and characterization of genes contributing to the oncogenic process, and determination of the chromosomal localization of novel genes.

3. My scientific Curriculum Vitae, including my list of publications, is attached to and forms part of this Declaration (Exhibit A).

4. I am familiar with a variety of techniques known in the art for detecting and quantifying the amplification of oncogenes in cancer, including the quantitative TaqMan PCR (i.e., "gene amplification") assay described in the above captioned patent application.

5. The TaqMan PCR assay is described, for example, in the following scientific publications: Higuchi *et al.*, Biotechnology 10:413-417 (1992) (Exhibit B); Livak *et al.*, PCR Methods Appl., 4:357-362 (1995) (Exhibit C) and Heid *et al.*, Genome Res. 6:986-994 (1996) (Exhibit D). Briefly, the assay is based on the principle that successful PCR yields a fluorescent signal due to Taq DNA polymerase-mediated exonuclease digestion of a fluorescently labeled oligonucleotide that is homologous to a sequence between two PCR primers. The extent of digestion depends directly on the amount of PCR, and can be quantified accurately by measuring the increment in fluorescence that results from decreased energy transfer. This is an extremely sensitive technique, which allows detection in the exponential phase of the PCR reaction and, as a result, leads to accurate determination of gene copy number.

6. The quantitative fluorescent TaqMan PCR assay has been extensively and successfully used to characterize genes involved in cancer development and progression. Amplification of protooncogenes has been studied in a variety of human tumors, and is widely considered as having etiological, diagnostic and prognostic significance. This use of the quantitative TaqMan PCR assay is exemplified by the following scientific publications: Pennica *et al.*, Proc. Natl. Acad. Sci. USA 95(25):14717-14722 (1998) (Exhibit E); Pitti *et al.*, Nature 396(6712):699-703 (1998) (Exhibit F) and Bieche *et al.*, Int. J. Cancer 78:661-666 (1998) (Exhibit G), the first two of which I am co-author. In particular, Pennica *et al.* have used the quantitative TaqMan PCR assay to study relative gene amplification of WISP and c-myc in various cell lines, colorectal tumors and normal mucosa. Pitti *et al.* studied the genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer, using the quantitative TaqMan PCR assay. Bieche *et al.* used the assay to study gene amplification in breast cancer.

7. It is my personal experience that the quantitative TaqMan PCR technique is technically sensitive enough to detect at least a 2-fold increase in gene copy number relative to control. It is further my considered scientific opinion that an at least 2-fold increase in gene copy number in a tumor tissue sample relative to a normal (i.e., non-tumor) sample is significant and useful in that the detected increase in gene copy number in the tumor sample relative to the normal sample serves as a basis for using relative gene copy number as quantitated by the TaqMan PCR technique as a diagnostic marker for the presence or absence of tumor in a tissue sample of unknown pathology. Accordingly, a gene identified as being amplified at least 2-fold by the quantitative TaqMan PCR assay in a tumor sample relative to a normal sample is useful as a marker for the diagnosis of cancer, for monitoring cancer development and/or for measuring the efficacy of cancer therapy.

8. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true. I declare that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Jan. 16, 2003
Date

Audrey Goddard
Audrey D. Goddard, Ph.D.

AUDREY D. GODDARD, Ph.D.

Genentech, Inc.
1 DNA Way
South San Francisco, CA, 94080
650.225.6429
goddarda@gene.com

110 Congo St.
San Francisco, CA, 94131
415.841.9154
415.819.2247 (mobile)
agoddard@pacbell.net

PROFESSIONAL EXPERIENCE

Genentech, Inc. **1993-present**
South San Francisco, CA

2001 - present **Senior Clinical Scientist**
Experimental Medicine / BioOncology, Medical Affairs

Responsibilities:
- *Companion diagnostic oncology products*
- *Acquisition of clinical samples from Genentech's clinical trials for translational research*
- *Translational research using clinical specimen and data for drug development and diagnostics*
- *Member of Development Science Review Committee, Diagnostic Oversight Team, 21 CFR Part 11 Subteam*

Interests:
- *Ethical and legal implications of experiments with clinical specimens and data*
- *Application of pharmacogenomics in clinical trials*

1998 - 2001 **Senior Scientist**
Head of the DNA Sequencing Laboratory, Molecular Biology Department, Research

Responsibilities:
- *Management of a laboratory of up to nineteen –including postdoctoral fellow, associate scientist, senior research associate and research assistants/associate levels*
- *Management of a $750K budget*
- *DNA sequencing core facility supporting a 350+ person research facility.*
- *DNA sequencing for high throughput gene discovery, - ESTs, cDNAs, and constructs*
- *Genomic sequence analysis and gene identification*
- *DNA sequence and primary protein analysis*

Research:
- *Chromosomal localization of novel genes*
- *Identification and characterization of genes contributing to the oncogenic process*
- *Identification and characterization of genes contributing to inflammatory diseases*
- *Design and development of schemes for high throughput genomic DNA sequence analysis*
- *Candidate gene prediction and evaluation*

1993 - 1998 Scientist
Head of the DNA Sequencing Laboratory, Molecular Biology Department, Research

Responsibilities
- *DNA sequencing core facility supporting a 350+ person research facility*
- *Assumed responsibility for a pre-existing team of five technicians and expanded the group into fifteen, introducing a level of middle management and additional areas of research*
- *Participated in the development of the basic plan for high throughput secreted protein discovery program – sequencing strategies, data analysis and tracking, database design*
- *High throughput EST and cDNA sequencing for new gene identification.*
- *Design and implementation of analysis tools required for high throughput gene identification.*
- *Chromosomal localization of genes encoding novel secreted proteins.*

Research:
- *Genomic sequence scanning for new gene discovery.*
- *Development of signal peptide selection methods.*
- *Evaluation of candidate disease genes.*
- *Growth hormone receptor gene SNPs in children with Idiopathic short stature*

Imperial Cancer Research Fund **1989-1992**
London, UK with Dr. Ellen Solomon

6/89 –12/92 Postdoctoral Fellow
- Cloning and characterization of the genes fused at the acute promyelocytic leukemia translocation breakpoints on chromosomes 17 and 15.
- Prepared a successfully funded European Union multi-center grant application

McMaster University **1983**
Hamilton, Ontario, Canada with Dr. G. D. Sweeney

5/83 – 8/83: NSERC Summer Student
- *In vitro* metabolism of β-naphthoflavone in C57Bl/6J and DBA mice

EDUCATION

Ph.D. "Phenotypic and genotypic effects of mutations in the human retinoblastoma gene." **Supervisor:** Dr. R. A. Phillips	University of Toronto Toronto, Ontario, Canada. Department of Medical Biophysics.	1989
Honours B.Sc "The *in vitro* metabolism of the cytochrome P-448 inducer β-naphthoflavone in C57BL/6J mice." **Supervisor:** Dr. G. D. Sweeney	McMaster University, Hamilton, Ontario, Canada. Department of Biochemistry	1983

ACADEMIC AWARDS

Imperial Cancer Research Fund Postdoctoral Fellowship	1989-1992
Medical Research Council Studentship	1983-1988
NSERC Undergraduate Summer Research Award	1983
Society of Chemical Industry Merit Award (Hons. Biochem.)	1983
Dr. Harry Lyman Hooker Scholarship	1981-1983
J.L.W. Gill Scholarship	1981-1982
Business and Professional Women's Club Scholarship	1980-1981
Wyerhauser Foundation Scholarship	1979-1980

INVITED PRESENTATIONS

Genentech's gene discovery pipeline: High throughput identification, cloning and characterization of novel genes. Functional Genomics: From Genome to Function, Litchfield Park, AZ, USA. October 2000

High throughput identification, cloning and characterization of novel genes. G2K:Back to Science, Advances in Genome Biology and Technology I. Marco Island, FL, USA. February 2000

Quality control in DNA Sequencing: The use of Phred and Phrap. Bay Area Sequencing Users Meeting, Berkeley, CA, USA. April 1999

High throughput secreted protein identification and cloning. Tenth International Genome Sequencing and Analysis Conference, Miami, FL, USA. September 1998

The evolution of DNA sequencing: The Genentech perspective. Bay Area Sequencing Users Meeting, Berkeley, CA, USA. May 1998

Partial Growth Hormone Insensitivity: The role of GH-receptor mutations in Idiopathic Short Stature. Tenth Annual National Cooperative Growth Study Investigators Meeting, San Francisco, CA, USA. October, 1996

Growth hormone (GH) receptor defects are present in selected children with non-GH-deficient short stature: A molecular basis for partial GH-insensitivity. 76th Annual Meeting of The Endocrine Society, Anaheim, CA, USA. June 1994

A previously uncharacterized gene, myl, is fused to the retinoic acid receptor alpha gene in acute promyelocytic leukemia. XV International Association for Comparative Research on Leukemia and Related Disease, Padua, Italy. October 1991

PATENTS

Goddard A, Godowski PJ, Gurney AL. NL2 Tie ligand homologue polypeptide. Patent Number: 6,455,496. Date of Patent: Sept. 24, 2002.

Goddard A, Godowski PJ and Gurney AL. NL3 Tie ligand homologue nucleic acids. Patent Number: 6,426,218. Date of Patent: July 30, 2002.

Godowski P, Gurney A, Hillan KJ, Botstein D, **Goddard A**, Roy M, Ferrara N, Tumas D, Schwall R. NL4 Tie ligand homologue nucleic acid. Patent Number: 6,4137,770. Date of Patent: July 2, 2002.

Ashkenazi A, Fong S, **Goddard A**, Gurney AL, Napier MA, Tumas D, Wood WI. Nucleic acid encoding A-33 related antigen poly peptides. Patent Number: 6,410,708. Date of Patent:: Jun. 25, 2002.

Botstein DA, Cohen RL, **Goddard AD**, Gurney AL, Hillan KJ, Lawrence DA, Levine AJ, Pennica D, Roy MA and Wood WI. WISP polypeptides and nucleic acids encoding same. Patent Number: 6,387,657. Date of Patent: May 14, 2002.

Goddard A, Godowski PJ and Gurney AL. Tie ligands. Patent Number: 6,372,491. Date of Patent: April 16, 2002.

Godowski PJ, Gurney AL, **Goddard A** and Hillan K. TIE ligand homologue antibody. Patent Number: 6,350,450. Date of Patent: Feb. 26, 2002.

Fong S, Ferrara N, **Goddard** A, Godowski PJ, Gurney AL, Hillan K and Williams PM. Tie receptor tyrosine kinase ligand homologues. Patent Number: 6,348,351. Date of Patent: Feb. 19, 2002.

Goddard A, Godowski PJ and Gurney AL. Ligand homologues. Patent Number: 6,348,350. Date of Patent: Feb. 19, 2002.

Attie KM, Carlsson LMS, Gesundheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 6,207,640. Date of Patent: March 27, 2001.

Fong S, Ferrara N, **Goddard A**, Godowski PJ, Gurney AL, Hillan K and Williams PM. Nucleic acids encoding NL-3. Patent Number: 6,074,873. Date of Patent: June 13, 2000

Attie K, Carlsson LMS, Gesunheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 5,824,642. Date of Patent: October 20, 1998

Attie K, Carlsson LMS, Gesunheit N and **Goddard A**. Treatment of partial growth hormone insensitivity syndrome. Patent Number: 5,646,113. Date of Patent: July 8, 1997

Multiple additional provisional applications filed

PUBLICATIONS

Seshasayee D, Dowd P, Gu Q, Erickson S, **Goddard AD** Comparative sequence analysis of the *HER2* locus in mouse and man. Manuscript in preparation.

Abuzzahab MJ, **Goddard A**, Grigorescu F, Lautier C, Smith RJ and Chernausek SD. Human IGF-1 receptor mutations resulting in pre- and post-natal growth retardation. Manuscript in preparation.

Aggarwal S, Xie, M-H, Foster J, Frantz G, Stinson J, Corpuz RT, Simmons L, Hillan K, Yansura DG, Vandlen RL, **Goddard AD** and Gurney AL. FHFR, a novel receptor for the fibroblast growth factors. Manuscript submitted.

Adams SH, Chui C, Schilbach SL, Yu XX, **Goddard AD**, Grimaldi JC, Lee J, Dowd P, Colman S., Lewin DA. (2001) BFIT, a unique acyl-CoA thioesterase induced in thermogenic brown adipose tissue: Cloning, organization of the human gene, and assessment of a potential link to obesity. *Biochemical Journal* **360**: 135-142.

Lee J. Ho WH. Maruoka M. Corpuz RT. Baldwin DT. Foster JS. **Goddard AD**. Yansura DG. Vandlen RL. Wood WI. Gurney AL. (2001) IL-17E, a novel proinflammatory ligand for the IL-17 receptor homolog IL-17Rh1. *Journal of Biological Chemistry* **276**(2): 1660-1664.

Xie M-H, Aggarwal S, Ho W-H, Foster J, Zhang Z, Stinson J, Wood WI, **Goddard AD** and Gurney AL. (2000) Interleukin (IL)-22, a novel human cytokine that signals through the interferon-receptor related proteins CRF2-4 and IL-22R. *Journal of Biological Chemistry* **275**: 31335-31339.

Weiss GA, Watanabe CK, Zhong A, **Goddard A** and Sidhu SS. (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA* **97**: 8950-8954.

Guo S, Yamaguchi Y, Schilbach S, Wada T.;Lee J, **Goddard A**, French D , Handa H, Rosenthal A. (2000) A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature* **408**: 366-369.

Yan M, Wang L-C, Hymowitz SG, Schilbach S, Lee J, **Goddard A**, de Vos AM, Gao WQ, Dixit VM. (2000) Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290**: 523-527.

Sehl PD, Tai JTN, Hillan KJ, Brown LA, **Goddard A**, Yang R, Jin H and Lowe DG. (2000) Application of cDNA microarrays in determining molecular phenotype in cardiac growth, development, and response to injury. *Circulation* **101**: 1990-1999.

Guo S, Brush J, Teraoka H, **Goddard A**, Wilson SW, Mullins MC and Rosenthal A. (1999) Development of noradrenergic neurons in the zebrafish hindbrain requires BMP, FGF8, and the homeodomain protein soulless/Phox2A. *Neuron* **24**: 555-566.

Stone D, Murone, M, Luoh, S, Ye W, Armanini P, Gurney A, Phillips HS, Brush, J, **Goddard A**, de Sauvage FJ and Rosenthal A. (1999) Characterization of the human suppressor of fused; a negative regulator of the zinc-finger transcription factor Gli. *J. Cell Sci.* **112**: 4437-4448.

Xie M-H, Holcomb I, Deuel B, Dowd P, Huang A, Vagts A, Foster J, Liang J, Brush J, Gu Q, Hillan K, **Goddard A** and Gurney, A.L. (1999) FGF-19, a novel fibroblast growth factor with unique specificity for FGFR4. *Cytokine* **11**: 729-735.

Yan M, Lee J, Schilbach S, **Goddard A** and Dixit V. (1999) mE10, a novel caspase recruitment domain-containing proapoptotic molecule. *J. Biol. Chem.* **274**(15): 10287-10292.

Gurney AL, Marsters SA, Huang RM, Pitti RM, Mark DT, Baldwin DT, Gray AM, Dowd P, Brush J, Heldens S, Schow P, **Goddard AD**, Wood WI, Baker KP, Godowski PJ and Ashkenazi A. (1999) Identification of a new member of the tumor necrosis factor family and its receptor, a human ortholog of mouse GITR. *Current Biology* **9**(4): 215-218.

Ridgway JBB, Ng E, Kern JA ,Lee J, Brush J, **Goddard A** and Carter P. (1999) Identification of a human anti-CD55 single-chain Fv by subtractive panning of a phage library using tumor and nontumor cell lines. *Cancer Research* 59: 2718-2723.

Pitti RM, Marsters SA, Lawrence DA, Roy M, Kischkel FC, Dowd P, Huang A, Donahue CJ, Sherwood SW, Baldwin DT, Godowski PJ, Wood WI, Gurney AL, Hillan KJ, Cohen RL, **Goddard AD**, Botstein D and Ashkenazi A. (1998) Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396**(6712): 699-703.

Pennica D, Swanson TA, Welsh JW, Roy MA, Lawrence DA, Lee J, Brush J, Taneyhill LA, Deuel B, Lew M, Watanabe C, Cohen RL, Melhem MF, Finley GG, Quirke P, **Goddard AD**, Hillan KJ, Gurney AL, Botstein D and Levine AJ. (1998) WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proc. Natl. Acad. Sci. USA*. **95**(25): 14717-14722.

Yang RB, Mark MR, Gray A, Huang A, Xie MH, Zhang M, **Goddard A**, Wood WI, Gurney AL and Godowski PJ. (1998) Toll-like receptor-2 mediates lipopolysaccharide-induced cellular signalling. *Nature* **395**(6699): 284-288.

Merchant AM, Zhu Z, Yuan JQ, **Goddard A**, Adams CW, Presta LG and Carter P. (1998) An efficient route to human bispecific IgG. *Nature Biotechnology* **16**(7): 677-681.

Marsters SA, Sheridan JP, Pitti RM, Brush J, **Goddard A** and Ashkenazi A. (1998) Identification of a ligand for the death-domain-containing receptor Apo3. *Current Biology* **8**(9): 525-528.

Xie J, Murone M, Luoh SM, Ryan A, Gu Q, Zhang C, Bonifas JM, Lam CW, Hynes M, **Goddard A,** Rosenthal A, Epstein EH Jr. and de Sauvage FJ. (1998) Activating Smoothened mutations in sporadic basal-cell carcinoma. *Nature*. **391**(6662): 90-92.

Marsters SA, Sheridan JP, Pitti RM, Huang A, Skubatch M, Baldwin D, Yuan J, Gurney A, **Goddard AD**, Godowski P and Ashkenazi A. (1997) A novel receptor for Apo2L/TRAIL contains a truncated death domain. *Current Biology*. **7**(12): 1003-1006.

Hynes M, Stone DM, Dowd M, Pitts-Meek S, **Goddard A**, Gurney A and Rosenthal A. (1997) Control of cell pattern in the neural tube by the zinc finger transcription factor *Gli-1*. *Neuron* **19**: 15–26.

Sheridan JP, Marsters SA, Pitti RM, Gurney A., Skubatch M, Baldwin D, Ramakrishnan L, Gray CL, Baker K, Wood WI, **Goddard AD**, Godowski P, and Ashkenazi A. (1997) Control of TRAIL-Induced Apoptosis by a Family of Signaling and Decoy Receptors. *Science* **277** (5327): 818-821.

Goddard AD, Dowd P, Chernausek S, Geffner M, Gertner J, Hintz R, Hopwood N, Kaplan S, Plotnick L, Rogol A, Rosenfield R, Saenger P, Mauras N, Hershkopf R, Angulo M and Attie, K. (1997) Partial growth hormone insensitivity: The role of growth hormone receptor mutations in idiopathic short stature. *J. Pediatr.* **131**: S51-55.

Klein RD, Sherman D, Ho WH, Stone D, Bennett GL, Moffat B, Vandlen R, Simmons L, Gu Q, Hongo JA, Devaux B, Poulsen K, Armanini M, Nozaki C, Asai N, **Goddard A**, Phillips H, Henderson CE, Takahashi M and Rosenthal A. (1997) A GPI-linked protein that interacts with Ret to form a candidate neurturin receptor. *Nature*. **387**(6634): 717-21.

Stone DM, Hynes M, Armanini M, Swanson TA, Gu Q, Johnson RL, Scott MP, Pennica D, **Goddard A**, Phillips H, Noll M, Hooper JE, de Sauvage F and Rosenthal A. (1996) The tumour-suppressor gene patched encodes a candidate receptor for Sonic hedgehog. *Nature* **384**(6605): 129-34.

Marsters SA, Sheridan JP, Donahue CJ, Pitti RM, Gray CL, **Goddard AD**, Bauer KD and Ashkenazi A. (1996) Apo-3, a new member of the tumor necrosis factor receptor family, contains a death domain and activates apoptosis and NF-kappa β. *Current Biology* **6**(12): 1669-76.

Rothe M, Xiong J, Shu HB, Williamson K, **Goddard A** and Goeddel DV. (1996) I-TRAF is a novel TRAF-interacting protein that regulates TRAF-mediated signal transduction. *Proc. Natl. Acad. Sci. USA* **93**: 8241-8246.

Yang M, Luoh SM, **Goddard A**, Reilly D, Henzel W and Bass S. (1996) The bglX gene located at 47.8 min on the Escherichia coli chromosome encodes a periplasmic beta-glucosidase. *Microbiology* **142**: 1659-65.

Goddard AD and Black DM. (1996) Familial Cancer in Molecular Endocrinology of Cancer. Waxman, J. Ed. Cambridge University Press, Cambridge UK, pp.187-215.

Treanor JJS, Goodman L, de Sauvage F, Stone DM, Poulson KT, Beck CD, Gray C, Armanini MP, Pollocks RA, Hefti F, Phillips HS, **Goddard A**, Moore MW, Buj-Bello A, Davis AM, Asai N, Takahashi M, Vandlen R, Henderson CE and Rosenthal A. (1996) Characterization of a receptor for GDNF. *Nature* **382**: 80-83.

Klein RD, Gu Q, **Goddard A** and Rosenthal A. (1996) Selection for genes encoding secreted proteins and receptors. *Proc. Natl. Acad. Sci. USA* **93**: 7108-7113.

Winslow JW, Moran P, Valverde J, Shih A, Yuan JQ, Wong SC, Tsai SP, **Goddard A**, Henzel WJ, Hefti F and Caras I. (1995) Cloning of AL-1, a ligand for an Eph-related tyrosine kinase receptor involved in axon bundle formation. *Neuron* **14**: 973-981.

Bennett BD, Zeigler FC, Gu Q, Fendly B, **Goddard AD**, Gillett N and Matthews W. (1995) Molecular cloning of a ligand for the EPH-related receptor protein-tyrosine kinase Htk. *Proc. Natl. Acad. Sci. USA* **92**: 1866-1870.

Huang X, Yuang J, **Goddard A**, Foulis A, James RF, Lernmark A, Pujol-Borrell R, Rabinovitch A, Somoza N and Stewart TA. (1995) Interferon expression in the pancreases of patients with type I diabetes. *Diabetes* **44**: 658-664.

Goddard AD, Yuan JQ, Fairbairn L, Dexter M, Borrow J, Kozak C and Solomon E. (1995) Cloning of the murine homolog of the leukemia-associated PML gene. *Mammalian Genome* **6**: 732-737.

Goddard AD, Covello R, Luoh SM, Clackson T, Attie KM, Gesundheit N, Rundle AC, Wells JA, Carlsson LMTI and The Growth Hormone Insensitivity Study Group. (1995) Mutations of the growth hormone receptor in children with idiopathic short stature. *N. Engl. J. Med.* **333**: 1093-1098.

Kuo SS, Moran P, Gripp J, Armanini M, Phillips HS, **Goddard A** and Caras IW. (1994) Identification and characterization of Batk, a predominantly brain-specific non-receptor protein tyrosine kinase related to Csk. *J. Neurosci. Res.* **38**: 705-715.

Mark MR, Scadden DT, Wang Z, Gu Q, **Goddard A** and Godowski PJ. (1994) Rse, a novel receptor-type tyrosine kinase with homology to Axl/Ufo, is expressed at high levels in the brain. *Journal of Biological Chemistry* **269**: 10720-10728.

Borrow J, Shipley J, Howe K, Kiely F, **Goddard A**, Sheer D, Srivastava A, Antony AC, Fioretos T, Mitelman F and Solomon E. (1994) Molecular analysis of simple variant translocations in acute promyelocytic leukemia. *Genes Chromosomes Cancer* **9**: 234-243.

Goddard AD and Solomon E. (1993) Genetics of Cancer. *Adv. Hum. Genet.* **21**: 321-376.

Borrow J, **Goddard AD,** Gibbons B, Katz F, Swirsky D, Fioretos T, Dube I, Winfield DA, Kingston J, Hagemeijer A, Rees JKH, Lister AT and Solomon E. (1992) Diagnosis of acute promyelocytic leukemia by RT-PCR: Detection of *PML-RARA* and *RARA-PML* fusion transcripts. *Br. J. Haematol.* **82**: 529-540.

Goddard AD, Borrow J and Solomon E. (1992) A previously uncharacterized gene, PML, is fused to the retinoic acid receptor alpha gene in acute promyelocytic leukemia. *Leukemia* **6 Suppl 3**: 117S–119S.

Zhu X, Dunn JM, **Goddard AD**, Squire JA, Becker A, Phillips RA and Gallie BL. (1992) Mechanisms of loss of heterozygosity in retinoblastoma. *Cytogenet. Cell. Genet.* **59**: 248-252.

Foulkes W, **Goddard A.** and Patel K. (1991) Retinoblastoma linked with Seascale [letter]. *British Med. J.* **302**: 409.

Goddard AD, Borrow J, Freemont PS and Solomon E. (1991) Characterization of a novel zinc finger gene disrupted by the t(15;17) in acute promyelocytic leukemia. *Science* **254**: 1371-1374.

Solomon E, Borrow J and **Goddard AD.** (1991) Chromosomal aberrations in cancer. *Science* **254**: 1153-1160.

Pajunen L, Jones TA, **Goddard A**, Sheer D, Solomon E, Pihlajaniemi T and Kivirikko KI. (1991) Regional assignment of the human gene coding for a multifunctional peptide (P4HB) acting as the β–subunit of prolyl-4-hydroxylase and the enzyme protein disulfide isomerase to 17q25. *Cytogenet. Cell. Genet.* **56**: 165-168.

Borrow J, Black DM, **Goddard AD**, Yagle MK, Frischauf A.-M and Solomon E. (1991) Construction and regional localization of a *Not*I linking library from human chromosome 17q. *Genomics* **10**: 477–480.

Borrow J, **Goddard AD**, Sheer D and Solomon E. (1990) Molecular analysis of acute promyelocytic leukemia breakpoint cluster region on chromosome 17. *Science* **249**: 1577-1580.

Myers JC, Jones TA, Pohjolainen E-R, Kadri AS, **Goddard AD**, Sheer D, Solomon E and Pihlajaniemi T. (1990) Molecular cloning of 5(IV) collagen and assignment of the gene to the region of the region of the X-chromosome containing the Alport Syndrome locus. *Am. J. Hum. Genet.* **46**: 1024-1033.

Gallie BL, Squire JA, **Goddard A**, Dunn JM, Canton M, Hinton D, Zhu X and Phillips RA. (1990) Mechanisms of oncogenesis in retinoblastoma. *Lab. Invest.* **62**: 394-408.

Goddard AD, Phillips RA, Greger V, Passarge E, Hopping W, Gallie BL and Horsthemke B. (1990) Use of the RB1 cDNA as a diagnostic probe in retinoblastoma families. *Clinical Genetics* **37**: 117-126.

Zhu XP, Dunn JM, Phillips RA, **Goddard AD**, Paton KE, Becker A and Gallie BL. (1989) Germline, but not somatic, mutations of the RB1 gene preferentially involve the paternal allele. *Nature* **340**: 312-314.

Gallie BL, Dunn JM, **Goddard A**, Becker A and Phillips RA. (1988) Identification of mutations in the putative retinoblastoma gene. In Molecular Biology of The Eye: Genes, Vision and Ocular Disease. UCLA Symposia on Molecular and Cellular Biology, New Series, Volume 88. J. Piatigorsky, T. Shinohara and P.S. Zelenka, Eds. Alan R. Liss, Inc., New York, 1988, pp. 427-436.

Goddard AD, Balakier H, Canton M, Dunn J, Squire J, Reyes E, Becker A, Phillips RA and Gallie BL. (1988) Infrequent genomic rearrangement and normal expression of the putative RB1 gene in retinoblastoma tumors. *Mol. Cell. Biol.* **8**: 2082-2088.

Squire J, Dunn J, **Goddard A**, Hoffman T, Musarella M, Willard HF, Becker AJ, Gallie BL and Phillips RA. (1986) Cloning of the esterase D gene: A polymorphic gene probe closely linked to the retinoblastoma locus on chromosome 13. *Proc. Natl. Acad. Sci.* USA **83**: 6573-6577.

Squire J, **Goddard AD**, Canton M, Becker A, Phillips RA and Gallie BL (1986) Tumour induction by the retinoblastoma mutation is independent of N-*myc* expression. *Nature* **322**: 555-557.

Goddard AD, Heddle JA, Gallie BL and Phillips RA. (1985) Radiation sensitivity of fibroblasts of bilateral retinoblastoma patients as determined by micronucleus induction *in vitro*. *Mutation Research* **152**: 31-38.

RESEARCH

SIMULTANEOUS AMPLIFICATION AND DETECTION OF SPECIFIC DNA SEQUENCES

Russell Higuchi*, Gavin Dollinger[1], P. Sean Walsh and Robert Griffith

Roche Molecular Systems, Inc., 1400 53rd St., Emeryville, CA 94608. [1]Chiron Corporation, 1400 53rd St., Emeryville, CA 94608. *Corresponding author.

We have enhanced the polymerase chain reaction (PCR) such that specific DNA sequences can be detected without opening the reaction tube. This enhancement requires the addition of ethidium bromide (EtBr) to a PCR. Since the fluorescence of EtBr increases in the presence of double-stranded (ds) DNA an increase in fluorescence in such a PCR indicates a positive amplification, which can be easily monitored externally. In fact, amplification can be continuously monitored in order to follow its progress. The ability to simultaneously amplify specific DNA sequences and detect the product of the amplification both simplifies and improves PCR and may facilitate its automation and more widespread use in the clinic or in other situations requiring high sample throughput.

Although the potential benefits of PCR[1] to clinical diagnostics are well known[2,3], it is still not widely used in this setting, even though it is four years since thermostable DNA polymerases[4] made PCR practical. Some of the reasons for its slow acceptance are high cost, lack of automation of pre- and post-PCR processing steps, and false positive results from carryover-contamination. The first two points are related in that labor is the largest contributor to cost at the present stage of PCR development. Most current assays require some form of "downstream" processing once thermocycling is done in order to determine whether the target DNA sequence was present and has amplified. These include DNA hybridization[5,6], gel electrophoresis with or without use of restriction digestion[7,8], HPLC[9], or capillary electrophoresis[10]. These methods are labor-intense, have low throughput, and are difficult to automate. The third point is also closely related to downstream processing. The handling of the PCR product in these downstream processes increases the chances that amplified DNA will spread through the typing lab, resulting in a risk of "carryover" false positives in subsequent testing[11].

These downstream processing steps would be eliminated if specific amplification and detection of amplified DNA took place simultaneously within an unopened reaction vessel. Assays in which such different processes take place without the need to separate reaction components have been termed "homogeneous". No truly homogeneous PCR assay has been demonstrated to date, although progress towards this end has been reported. Chehab, et al.[12], developed a PCR product detection scheme using fluorescent primers that resulted in a fluorescent PCR product. Allele-specific primers, each with different fluorescent tags, were used to indicate the genotype of the DNA. However, the unincorporated primers must still be removed in a downstream process in order to visualize the result. Recently, Holland, et al.[13], developed an assay in which the endogenous 5′ exonuclease assay of *Taq* DNA polymerase was exploited to cleave a labeled oligonucleotide probe. The probe would only cleave if PCR amplification had produced its complementary sequence. In order to detect the cleavage products, however, a subsequent process is again needed.

We have developed a truly homogeneous assay for PCR and PCR product detection based upon the greatly increased fluorescence that ethidium bromide and other DNA binding dyes exhibit when they are bound to ds-DNA[14–16]. As outlined in Figure 1, a prototypic PCR

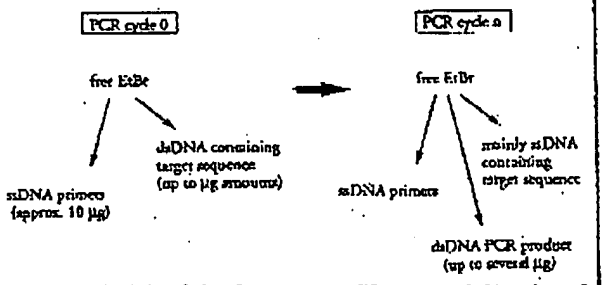


FIGURE 1 Principle of simultaneous amplification and detection of PCR product. The components of a PCR containing EtBr that are fluorescent are listed—EtBr itself, EtBr bound to either ssDNA or dsDNA. There is a large fluorescence enhancement when EtBr is bound to DNA and binding is greatly enhanced when DNA is double-stranded. After sufficient (n) cycles of PCR, the net increase in dsDNA results in additional EtBr binding, and a net increase in total fluorescence.

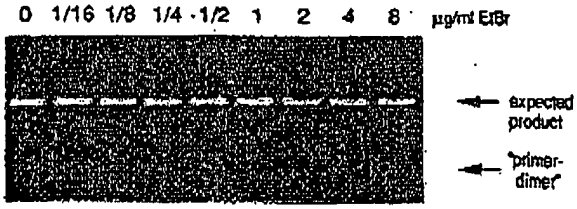


FIGURE 2 Gel electrophoresis of PCR amplification products of the human, nuclear gene, HLA DQα, made in the presence of increasing amounts of EtBr (up to 8 µg/ml). The presence of EtBr has no obvious effect on the yield or specificity of amplification.

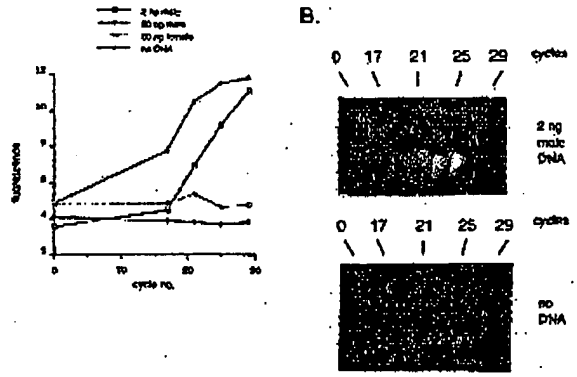


FIGURE 3 (A) Fluorescence measurements from PCRs that contain 0.5 µg/ml EtBr and that are specific for Y-chromosome repeat sequences. Five replicate PCRs were begun containing each of the DNAs specified. At each indicated cycle, one of the five replicate PCRs for each DNA was removed from thermocycling and its fluorescence measured. Units of fluorescence are arbitrary. (B) UV photography of PCR tubes (0.5 ml Eppendorf-style, polypropylene micro-centrifuge tubes) containing reactions, those starting from 2 ng male DNA and control reactions without any DNA, from (A).

begins with primers that are single-stranded DNA (ssDNA), dNTPs, and DNA polymerase. An amount of dsDNA containing the target sequence (target DNA) is also typically present. This amount can vary, depending on the application, from single-cell amounts of DNA[17] to micrograms per PCR[18]. If EtBr is present, the reagents that will fluoresce, in order of increasing fluorescence, are free EtBr itself, and EtBr bound to the single-stranded DNA primers and to the double-stranded target DNA (by its intercalation between the stacked bases of the DNA double-helix). After the first denaturation cycle, target DNA will be largely single-stranded. After a PCR is completed, the most significant change is the increase in the amount of dsDNA (the PCR product itself) of up to several micrograms. Formerly free EtBr is bound to the additional dsDNA, resulting in an increase in fluorescence. There is also some decrease in the amount of ssDNA primer, but because the binding of EtBr to ssDNA is much less than to dsDNA, the effect of this change on the total fluorescence of the sample is small. The fluorescence increase can be measured by directing excitation illumination through the walls of the amplification vessel before and after, or even continuously during, thermocycling.

RESULTS

PCR in the presence of EtBr. In order to assess the affect of EtBr in PCR, amplifications of the human HLA DQα gene[19] were performed with the dye present at concentrations from 0.06 to 8.0 µg/ml (a typical concentration of EtBr used in staining of nucleic acids following gel electrophoresis is 0.5 µg/ml). As shown in Figure 2, gel electrophoresis revealed little or no difference in the yield or quality of the amplification product whether EtBr was absent or present at any of these concentrations, indicating that EtBr does not inhibit PCR.

Detection of human Y-chromosome specific sequences. Sequence-specific, fluorescence enhancement of EtBr as a result of PCR was demonstrated in a series of amplifications containing 0.5 µg/ml EtBr and primers specific to repeat DNA sequences found on the human Y-chromosome[20]. These PCRs initially contained either 60 ng male, 60 ng female, 2 ng male human or no DNA. Five replicate PCRs were begun for each DNA. After 0, 17, 21, 24 and 29 cycles of thermocycling, a PCR for each DNA was removed from the thermocycler, and its fluorescence measured in a spectrofluorometer and plotted vs. amplification cycle number (Fig. 3A). The shape of this curve reflects the fact that by the time an increase in fluorescence can be detected, the increase in DNA is becoming linear and not exponential with cycle number. As shown, the fluorescence increased about three-fold over the background fluorescence for the PCRs containing human male DNA, but did not significantly increase for negative control PCRs, which contained either no DNA or human female DNA. The more male DNA present to begin with—60 ng versus 2 ng—the fewer cycles were needed to give a detectable increase in fluorescence. Gel electrophoresis on the products of these amplifications showed that DNA fragments of the expected size were made in the male DNA containing reactions and that little DNA synthesis took place in the control samples.

In addition, the increase in fluorescence was visualized by simply laying the completed, unopened PCRs on a UV transilluminator and photographing them through a red filter. This is shown in figure 3B for the reactions that began with 2 ng male DNA and those with no DNA.

Detection of specific alleles of the human β-globin gene. In order to demonstrate that this approach has adequate specificity to allow genetic screening, a detection of the sickle-cell anemia mutation was performed. Figure 4 shows the fluorescence from completed amplifications containing EtBr (0.5 µg/ml) as detected by photography of the reaction tubes on a UV transilluminator. These reactions were performed using primers specific for either the wild-type or sickle-cell mutation of the human β-globin gene[21]. The specificity for each allele is imparted by placing the sickle-mutation site at the terminal 3' nucleotide of one primer. By using an appropriate primer annealing temperature, primer extension—and thus amplification—can take place only if the 3' nucleotide of the primer is complementary to the β-globin allele present[21,22].

Each pair of amplifications shown in Figure 4 consists of a reaction with either the wild-type allele specific (left tube) or sickle-allele specific (right tube) primers. Three different DNAs were typed: DNA from a homozygous, wild-type β-globin individual (AA); from a heterozygous sickle β-globin individual (AS); and from a homozygous sickle β-globin individual (SS). Each DNA (50 ng genomic DNA to start each PCR) was analyzed in triplicate (3 pairs

of reactions each). The DNA type was reflected in the relative fluorescence intensities in each pair of completed amplifications. There was a significant increase in fluorescence only where a β-globin allele DNA matched the primer set. When measured on a spectrofluorometer (data not shown), this fluorescence was about three times that present in a PCR where both β-globin alleles were mismatched to the primer set. Gel electrophoresis (not shown) established that this increase in fluorescence was due to the synthesis of nearly a microgram of a DNA fragment of the expected size for β-globin. There was little synthesis of dsDNA in reactions in which the allele-specific primer was mismatched to both alleles.

Continuous monitoring of a PCR. Using a fiber optic device, it is possible to direct excitation illumination from a spectrofluorometer to a PCR undergoing thermocycling and to return its fluorescence to the spectrofluorometer. The fluorescence readout of such an arrangement, directed at an EtBr-containing amplification of Y-chromosome specific sequences from 25 ng of human male DNA, is shown in Figure 5. The readout from a control PCR with no target DNA is also shown. Thirty cycles of PCR were monitored for each.

The fluorescence trace as a function of time clearly shows the effect of the thermocycling. Fluorescence intensity rises and falls inversely with temperature. The fluorescence intensity is minimum at the denaturation temperature (94°C) and maximum at the annealing/extension temperature (50°C). In the negative-control PCR, these fluorescence maxima and minima do not change significantly over the thirty thermocycles, indicating that there is little dsDNA synthesis without the appropriate target DNA, and there is little if any bleaching of EtBr during the continuous illumination of the sample.

In the PCR containing male DNA, the fluorescence maxima at the annealing/extension temperature begin to increase at about 4000 seconds of thermocycling, and continue to increase with time, indicating that dsDNA is being produced at a detectable level. Note that the fluorescence minima at the denaturation temperature do not significantly increase, presumably because at this temperature there is no dsDNA for EtBr to bind. Thus the course of the amplification is followed by tracking the fluorescence increase at the annealing temperature. Analysis of the products of these two amplifications by gel electrophoresis showed a DNA fragment of the expected size for the male DNA containing sample and no detectable DNA synthesis for the control sample.

DISCUSSION

Downstream processes such as hybridization to a sequence-specific probe can enhance the specificity of DNA detection by PCR. The elimination of these processes means that the specificity of this homogeneous assay depends solely on that of PCR. In the case of sickle-cell disease, we have shown that PCR alone has sufficient DNA sequence specificity to permit genetic screening. Using appropriate amplification conditions, there is little non-specific production of dsDNA in the absence of the appropriate target allele.

The specificity required to detect pathogens can be more or less than that required to do genetic screening, depending on the number of pathogens in the sample and the amount of other DNA that must be taken with the sample. A difficult target is HIV, which requires detection of a viral genome that can be at the level of a few copies per thousands of host cells[6]. Compared with genetic screening, which is performed on cells containing at least one copy of the target sequence, HIV detection requires both more specificity and the input of more total

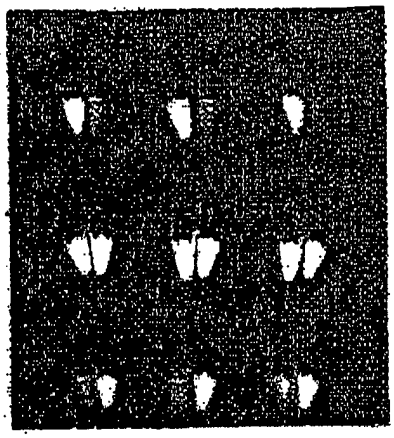


FIGURE 4 UV photography of PCR tubes containing amplifications using EtBr that are specific to wild-type (A) or sickle (S) alleles of the human β-globin gene. The left of each pair of tubes contains allele-specific primers to the wild-type alleles, the right tube primers to the sickle allele. The photograph was taken after 30 cycles of PCR, and the input DNAs and the alleles they contain are indicated. Fifty ng of DNA was used to begin PCR. Typing was done in triplicate (3 pairs of PCRs) for each input DNA.

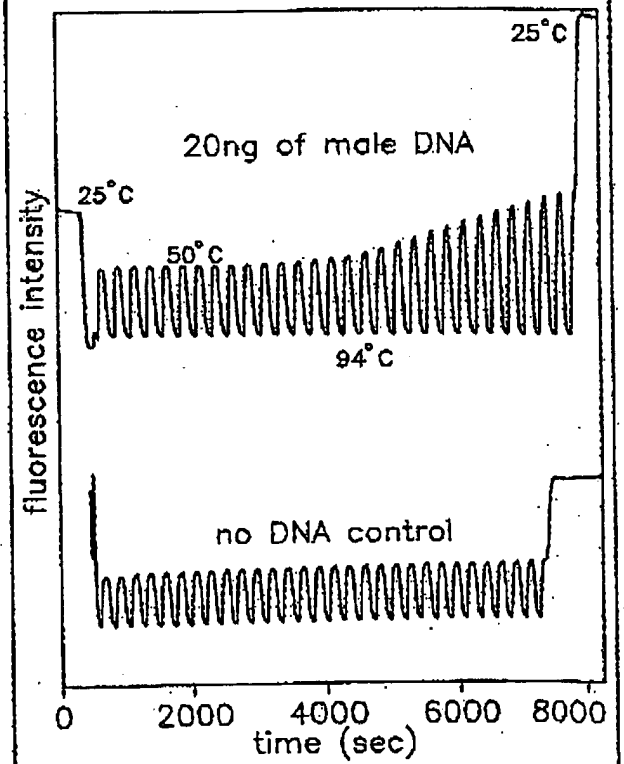


FIGURE 5 Continuous, real-time monitoring of a PCR. A fiber optic was used to carry excitation light to a PCR in progress and also emitted light back to a fluorometer (see Experimental Protocol). Amplification using human male-DNA specific primers in a PCR starting with 20 ng of human male DNA (top), or in a control PCR without DNA (bottom), were monitored. Thirty cycles of PCR were followed for each. The temperature cycled between 94°C (denaturation) and 50°C (annealing and extension). Note in the male DNA PCR, the cycle (time) dependent increase in fluorescence at the annealing/extension temperature.

DNA—up to microgram amounts—in order to have sufficient numbers of target sequences. This large amount of starting DNA in an amplification significantly increases the background fluorescence over which any additional fluorescence produced by PCR must be detected. An additional complication that occurs with targets in low copy-number is the formation of the "primer-dimer" artifact. This is the result of the extension of one primer using the other primer as a template. Although this occurs infrequently, once it occurs the extension product is a substrate for PCR amplification, and can compete with true PCR targets if those targets are rare. The primer-dimer product is of course dsDNA and thus is a potential source of false signal in this homogeneous assay.

To increase PCR specificity and reduce the effect of primer-dimer amplification, we are investigating a number of approaches, including the use of nested-primer amplifications that take place in a single tube[5], and the "hot-start", in which nonspecific amplification is reduced by raising the temperature of the reaction before DNA synthesis begins[23]. Preliminary results using these approaches suggest that primer-dimer is effectively reduced and it is possible to detect the increase in EtBr fluorescence in a PCR instigated by a single HIV genome in a background of 10^5 cells. With larger numbers of cells, the background fluorescence contributed by genomic DNA becomes problematic. To reduce this background, it may be possible to use sequence-specific DNA-binding dyes that can be made to preferentially bind PCR product over genomic DNA by incorporating the dye-binding DNA sequence into the PCR product through a 5′ "add-on" to the oligonucleotide primer[24].

We have shown that the detection of fluorescence generated by an EtBr-containing PCR is straightforward, both once PCR is completed and continuously during thermocycling. The ease with which automation of specific DNA detection can be accomplished is the most promising aspect of this assay. The fluorescence analysis of completed PCRs is already possible with existing instrumentation in 96-well format[25]. In this format, the fluorescence in each PCR can be quantitated before, after, and even at selected points during thermocycling by moving the rack of PCRs to a 96-microwell plate fluorescence reader[26].

The instrumentation necessary to continuously monitor multiple PCRs simultaneously is also simple in principle. A direct extension of the apparatus used here is to have multiple fiberoptics transmit the excitation light and fluorescent emissions to and from multiple PCRs. The ability to monitor multiple PCRs continuously may allow quantitation of target DNA copy number. Figure 3 shows that the larger the amount of starting target DNA, the sooner during PCR a fluorescence increase is detected. Preliminary experiments (Higuchi and Dollinger, manuscript in preparation) with continuous monitoring have shown a sensitivity to two-fold differences in initial target DNA concentration.

Conversely, if the number of target molecules is known—as it can be in genetic screening—continuous monitoring may provide a means of detecting false positive and false negative results. With a known number of target molecules, a true positive would exhibit detectable fluorescence by a predictable number of cycles of PCR. Increases in fluorescence detected before or after that cycle would indicate potential artifacts. False negative results due to, for example, inhibition of DNA polymerase, may be detected by including within each PCR an inefficiently amplifying marker. This marker results in a fluorescence increase only after a large number of cycles—many more than are necessary to detect a true positive. If a sample fails to have a fluorescence increase after this many cycles, inhibition may be suspected. Since, in this assay, conclusions are drawn based on the presence or absence of fluorescence signal alone, such controls may be important. In any event, before any test based on this principle is ready for the clinic, an assessment of its false positive/false negative rates will need to be obtained using a large number of known samples.

In summary, the inclusion in PCR of dyes whose fluorescence is enhanced upon binding dsDNA makes it possible to detect specific DNA amplification from outside the PCR tube. In the future, instruments based upon this principle may facilitate the more widespread use of PCR in applications that demand the high throughput of samples.

EXPERIMENTAL PROTOCOL

Human HLA-DQα gene amplifications containing EtBr. PCRs were set up in 100 µl volumes containing 10 mM Tris-HCl, pH 8.3; 50 mM KCl; 4 mM $MgCl_2$; 2.5 units of *Taq* DNA polymerase (Perkin-Elmer Cetus, Norwalk, CT); 20 pmole each of human HLA-DQα gene specific oligonucleotide primers GH26 and GH27[19] and approximately 10^3 copies of DQα PCR product diluted from a previous reaction. Ethidium bromide (EtBr; Sigma) was used at the concentrations indicated in Figure 2. Thermocycling proceeded for 20 cycles in a model 480 thermocycler (Perkin-Elmer Cetus, Norwalk, CT) using a "step-cycle" program of 94°C for 1 min. denaturation and 60°C for 30 sec. annealing and 72°C for 30 sec. extension.

Y-chromosome specific PCR. PCRs (100 µl total reaction volume) containing 0.5 µg/ml EtBr were prepared as described for HLA-DQα, except with different primers and target DNAs. These PCRs contained 15 pmole each male DNA-specific primers Y1.1 and Y1.2[20], and either 60 ng male, 60 ng female, 2 ng male, or no human DNA. Thermocycling was 94°C for 1 min. and 60°C for 1 min using a "step-cycle" program. The number of cycles for a sample were as indicated in Figure 3. Fluorescence measurement is described below.

Allele-specific, human β-globin gene PCR. Amplifications of 100 µl volume using 0.5 µg/ml of EtBr were prepared as described for HLA-DQα above except with different primers and target DNAs. These PCRs contained either primer pair HGP2/HβI4A (wild-type globin specific primers) or HGP2/HβI4S (sickle-globin specific primers) at 10 pmole each primer per PCR. These primers were developed by Wu et al.[21]. Three different target DNAs were used in separate amplifications—50 ng each of human DNA that was homozygous for the sickle trait (SS), DNA that was heterozygous for the sickle trait (AS), or DNA that was homozygous for the w.t. globin (AA). Thermocycling was for 30 cycles at 94°C for 1 min. and 55°C for 1 min. using a "step-cycle" program. An annealing temperature of 55°C had been shown by Wu et al.[21] to provide allele-specific amplification. Completed PCRs were photographed through a red filter (Wratten 23A) after placing the reaction tubes atop a model TM-36 transilluminator (UV-products San Gabriel, CA).

Fluorescence measurement. Fluorescence measurements were made on PCRs containing EtBr in a Fluorolog-2 fluorometer (SPEX, Edison, NJ). Excitation was at the 500 nm band with about 2 nm bandwidth with a GG 435 nm cut-off filter (Melles Grist, Inc., Irvine, CA) to exclude second-order light. Emitted light was detected at 570 nm with a bandwidth of about 7 nm. An OG 530 nm cut-off filter was used to remove the excitation light.

Continuous fluorescence monitoring of PCR. Continuous monitoring of a PCR in progress was accomplished using the spectrofluorometer and settings described above as well as a fiberoptic accessory (SPEX cat. no. 1950) to both send excitation light to, and receive emitted light from, a PCR placed in a well of a model 480 thermocycler (Perkin-Elmer Cetus). The probe end of the fiberoptic cable was attached with "5 minute-epoxy" to the open top of a PCR tube (a 0.5 ml polypropylene centrifuge tube with its cap removed) effectively sealing it. The exposed top of the PCR tube and the end of the fiberoptic cable were shielded from room light and the room lights were kept dimmed during each run. The monitored PCR was an amplification of Y-chromosome-specific repeat sequences as described above, except using an annealing/extension temperature of 50°C. The reaction was covered with mineral oil (2 drops) to prevent evaporation. Thermocycling and fluorescence measurement were started simultaneously. A time-base scan with a 10 second integration time

was used and the emission signal was ratioed to the excitation signal to control for changes in light-source intensity. Data were collected using the dm3000f, version 2.5 (SPEX) data system.

Acknowledgments

We thank Bob Jones for help with the spectrofluormetric measurements and Heatherbell Fong for editing this manuscript.

References

1. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. 1986. Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction. CSHSQB 51:263–273.
2. White, T. J., Arnheim, N. and Erlich, H. A. 1989. The polymerase chain reaction. Trends Genet. 5:185–189.
3. Erlich, H. A., Gelfand, D. and Sninsky, J. J. 1991. Recent advances in the polymerase chain reaction. Science 252:1643–1651.
4. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491.
5. Saiki, R. K., Walsh, P. S., Levenson, C. H. and Erlich, H. A. 1989. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc. Natl. Acad. Sci. USA 86:6230–6234.
6. Kwok, S. Y., Mack, D. H., Mullis, K. B., Poiesz, B. J., Ehrlich, G. D., Blair, D. and Friedman-Kien, A. S. 1987. Identification of human immunodeficiency virus sequences by using *in vitro* enzymatic amplification and oligomer cleavage detection. J. Virol. 61:1690–1694.
7. Chehab, F. F., Doherty, M., Cai, S. P., Kan, Y. W., Cooper, S. and Rubin, E. M. 1987. Detection of sickle cell anemia and thalassemias. Nature 329:293–294.
8. Horn, G. T., Richards, B. and Klinger, K. W. 1989. Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. Nuc. Acids Res. 16:2140.
9. Katz, E. D. and Dong, M. W. 1990. Rapid analysis and purification of polymerase chain reaction products by high-performance liquid chromatography. Biotechniques 8:546–555.
10. Heiger, D. N., Cohen, A. S. and Karger, B. L. 1990. Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using continuous and pulsed electric fields. J. Chromatogr. 516:33–48.
11. Kwok, S. Y. and Higuchi, R. G. 1989. Avoiding false positives with PCR. Nature 339:237–238.
12. Chehab, F. F. and Kan, Y. W. 1989. Detection of specific DNA sequences by fluorescence amplification: a color complementation assay. Proc. Natl. Acad. Sci. USA 86:9178–9182.
13. Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5′ to 3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc. Natl. Acad. Sci. USA 88:7276–7280.
14. Markovits, J., Roques, B. P. and Le Pecq, J. B. 1979. Ethidium dimer: a new reagent for the fluorimetric determination of nucleic acids. Anal. Biochem. 94:259–264.
15. Kapuscinski, J. and Szer, W. 1979. Interactions of 4′,6-diamidine-2-phenylindole with synthetic polynucleotides. Nuc. Acids Res. 6:3519–3534.
16. Searle, M. S. and Embrey, K. J. 1990. Sequence-specific interaction of Hoescht 33258 with the minor groove of an adenine-tract DNA duplex studied in solution by ^{1}H NMR spectroscopy. Nuc. Acids Res. 18:3753–3762.
17. Li, H. H., Gyllensten, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A. and Arnheim, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature 335:414–417.
18. Abbott, M. A., Poiesz, B. J., Byrne, B. C., Kwok, S. Y., Sninsky, J. J. and Erlich, H. A. 1988. Enzymatic gene amplification: qualitative and quantitative methods for detecting proviral DNA amplified *in vitro*. J. Infect. Dis. 158:1158.
19. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1986. Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. Nature 324:163–166.
20. Kogan, S. C., Doherty, M. and Gitschier, J. 1987. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. N. Engl. J. Med. 317:985–990.
21. Wu, D. Y., Ugozzoli, L., Pal, B. K. and Wallace, R. B. 1989. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. Proc. Natl. Acad. Sci. USA 86:2757–2760.
22. Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. Nuc. Acids Res. 18:999–1005.
23. Chou, Q., Russell, M., Birch, D., Raymond, J. and Bloch, W. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Submitted.
24. Higuchi, R. 1989. Using PCR to engineer DNA, p. 61–70. *In:* PCR Technology. H. A. Erlich (Ed.). Stockton Press, New York, N.Y.
25. Hall, L., Atwood, J. G., DiCesare, J., Katz, E., Pizzaza, E., Williams, J. F. and Woudenberg, T. 1991. A high-performance system for automation of the polymerase chain reaction. Biotechniques 10:102–103, 106–112.
26. Tumosa, N. and Kahan, L. 1989. Fluorescent EIA screening of monoclonal antibodies to cell surface antigens. J. Immun. Meth. 116:59–63.

GENENTECH, INC.
1 DNA Way
South San Francisco, CA 94080 USA
Phone: (650) 225-1000

FAX: (650) 952-9881

FACSIMILE TRANSMITTAL

Date: 19 July 2004

To: **Anna Barry**
Heller Ehrman

Re: Higuchi reference

Fax No: 324-6638

From: Patty Tobin, Assistant to Elizabeth M. Barnes, Ph.D.
Genentech, Inc. Legal Department

Number of Pages including this cover sheet: 6

RESEARCH

SIMULTANEOUS AMPLIFICATION AND DETECTION OF SPECIFIC DNA SEQUENCES

Russell Higuchi*, Gavin Dollinger[1], P. Sean Walsh and Robert Griffith

Roche Molecular Systems, Inc., 1400 53rd St., Emeryville, CA 94608. [1]Chiron Corporation, 1400 53rd St., Emeryville, CA 94608. *Corresponding author.

We have enhanced the polymerase chain reaction (PCR) such that specific DNA sequences can be detected without opening the reaction tube. This enhancement requires the addition of ethidium bromide (EtBr) to a PCR. Since the fluorescence of EtBr increases in the presence of double-stranded (ds) DNA an increase in fluorescence in such a PCR indicates a positive amplification, which can be easily monitored externally. In fact, amplification can be continuously monitored in order to follow its progress. The ability to simultaneously amplify specific DNA sequences and detect the product of the amplification both simplifies and improves PCR and may facilitate its automation and more widespread use in the clinic or in other situations requiring high sample throughput.

Although the potential benefits of PCR[1] to clinical diagnostics are well known[2,3], it is still not widely used in this setting, even though it is four years since thermostable DNA polymerases[4] made PCR practical. Some of the reasons for its slow acceptance are high cost, lack of automation of pre- and post-PCR processing steps, and false positive results from carryover-contamination. The first two points are related in that labor is the largest contributor to cost at the present stage of PCR development. Most current assays require some form of "downstream" processing once thermocycling is done in order to determine whether the target DNA sequence was present and has amplified. These include DNA hybridization[5,6], gel electrophoresis with or without use of restriction digestion[7,8], HPLC[9], or capillary electrophoresis[10]. These methods are labor-intense, have low throughput, and are difficult to automate. The third point is also closely related to downstream processing. The handling of the PCR product in these downstream processes increases the chances that amplified DNA will spread through the typing lab, resulting in a risk of "carryover" false positives in subsequent testing[11].

These downstream processing steps would be eliminated if specific amplification and detection of amplified DNA took place simultaneously within an unopened reaction vessel. Assays in which such different processes take place without the need to separate reaction components have been termed "homogeneous". No truly homogeneous PCR assay has been demonstrated to date, although progress towards this end has been reported. Chehab, et al.[12], developed a PCR product detection scheme using fluorescent primers that resulted in a fluorescent PCR product. Allele-specific primers, each with different fluorescent tags, were used to indicate the genotype of the DNA. However, the unincorporated primers must still be removed in a downstream process in order to visualize the result. Recently, Holland, et al.[13], developed an assay in which the endogenous 5′ exonuclease assay of *Taq* DNA polymerase was exploited to cleave a labeled oligonucleotide probe. The probe would only cleave if PCR amplification had produced its complementary sequence. In order to detect the cleavage products, however, a subsequent process is again needed.

We have developed a truly homogeneous assay for PCR and PCR product detection based upon the greatly increased fluorescence that ethidium bromide and other DNA binding dyes exhibit when they are bound to dsDNA[14–16]. As outlined in Figure 1, a prototypic PCR

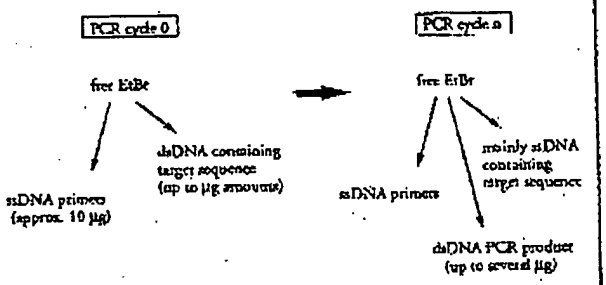


FIGURE 1 Principle of simultaneous amplification and detection of PCR product. The components of a PCR containing EtBr that are fluorescent are listed—EtBr itself, EtBr bound to either ssDNA or dsDNA. There is a large fluorescence enhancement when EtBr is bound to DNA and binding is greatly enhanced when DNA is double-stranded. After sufficient (n) cycles of PCR, the net increase in dsDNA results in additional EtBr binding, and a net increase in total fluorescence.

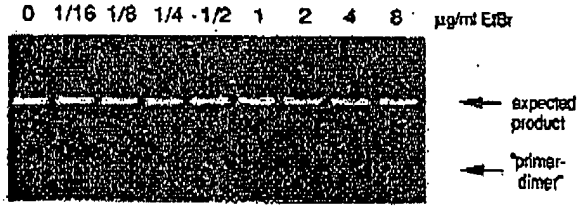


FIGURE 2 Gel electrophoresis of PCR amplification products of the human, nuclear gene, HLA DQα, made in the presence of increasing amounts of EtBr (up to 8 µg/ml). The presence of EtBr has no obvious effect on the yield or specificity of amplification.

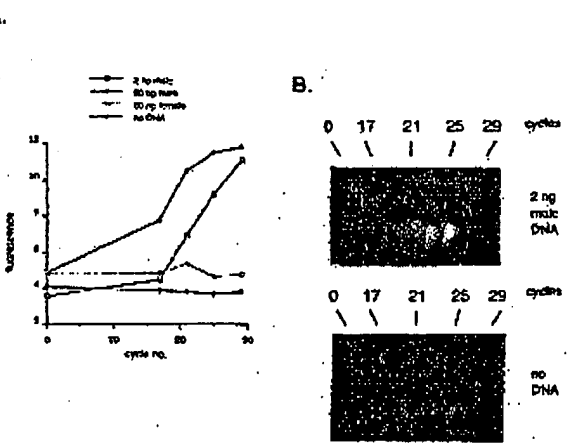


FIGURE 3 (A) Fluorescence measurements from PCRs that contain 0.5 µg/ml EtBr and that are specific for Y-chromosome repeat sequences. Five replicate PCRs were begun containing each of the DNAs specified. At each indicated cycle, one of the five replicate PCRs for each DNA was removed from thermocycling and its fluorescence measured. Units of fluorescence are arbitrary. (B) UV photography of PCR tubes (0.5 ml Eppendorf-style, polypropylene micro-centrifuge tubes) containing reactions, those starting from 2 ng male DNA and control reactions without any DNA, from (A).

begins with primers that are single-stranded DNA (ssDNA), dNTPs, and DNA polymerase. An amount of dsDNA containing the target sequence (target DNA) is also typically present. This amount can vary, depending on the application, from single-cell amounts of DNA[17] to micrograms per PCR[18]. If EtBr is present, the reagents that will fluoresce, in order of increasing fluorescence, are free EtBr itself, and EtBr bound to the single-stranded DNA primers and to the double-stranded target DNA (by its intercalation between the stacked bases of the DNA double-helix). After the first denaturation cycle, target DNA will be largely single-stranded. After a PCR is completed, the most significant change is the increase in the amount of dsDNA (the PCR product itself) of up to several micrograms. Formerly free EtBr is bound to the additional dsDNA, resulting in an increase in fluorescence. There is also some decrease in the amount of ssDNA primer, but because the binding of EtBr to ssDNA is much less than to dsDNA, the effect of this change on the total fluorescence of the sample is small. The fluorescence increase can be measured by directing excitation illumination through the walls of the amplification vessel before and after, or even continuously during, thermocycling.

RESULTS

PCR in the presence of EtBr. In order to assess the affect of EtBr in PCR, amplifications of the human HLA DQα gene[19] were performed with the dye present at concentrations from 0.06 to 8.0 µg/ml (a typical concentration of EtBr used in staining of nucleic acids following gel electrophoresis is 0.5 µg/ml). As shown in Figure 2, gel electrophoresis revealed little or no difference in the yield or quality of the amplification product whether EtBr was absent or present at any of these concentrations, indicating that EtBr does not inhibit PCR.

Detection of human Y-chromosome specific sequences. Sequence-specific, fluorescence enhancement of EtBr as a result of PCR was demonstrated in a series of amplifications containing 0.5 µg/ml EtBr and primers specific to repeat DNA sequences found on the human Y-chromosome[20]. These PCRs initially contained either 60 ng male, 60 ng female, 2 ng male human or no DNA. Five replicate PCRs were begun for each DNA. After 0, 17, 21, 24 and 29 cycles of thermocycling, a PCR for each DNA was removed from the thermocycler, and its fluorescence measured in a spectrofluorometer and plotted vs. amplification cycle number (Fig. 3A). The shape of this curve reflects the fact that by the time an increase in fluorescence can be detected, the increase in DNA is becoming linear and not exponential with cycle number. As shown, the fluorescence increased about three-fold over the background fluorescence for the PCRs containing human male DNA, but did not significantly increase for negative control PCRs, which contained either no DNA or human female DNA. The more male DNA present to begin with—60 ng versus 2 ng—the fewer cycles were needed to give a detectable increase in fluorescence. Gel electrophoresis on the products of these amplifications showed that DNA fragments of the expected size were made in the male DNA containing reactions and that little DNA synthesis took place in the control samples.

In addition, the increase in fluorescence was visualized by simply laying the completed, unopened PCRs on a UV transilluminator and photographing them through a red filter. This is shown in figure 3B for the reactions that began with 2 ng male DNA and those with no DNA.

Detection of specific alleles of the human β-globin gene. In order to demonstrate that this approach has adequate specificity to allow genetic screening, a detection of the sickle-cell anemia mutation was performed. Figure 4 shows the fluorescence from completed amplifications containing EtBr (0.5 µg/ml) as detected by photography of the reaction tubes on a UV transilluminator. These reactions were performed using primers specific for either the wild-type or sickle-cell mutation of the human β-globin gene[21]. The specificity for each allele is imparted by placing the sickle-mutation site at the terminal 3′ nucleotide of one primer. By using an appropriate primer annealing temperature, primer extension—and thus amplification—can take place only if the 3′ nucleotide of the primer is complementary to the β-globin allele present[21,22].

Each pair of amplifications shown in Figure 4 consists of a reaction with either the wild-type allele specific (left tube) or sickle-allele specific (right tube) primers. Three different DNAs were typed: DNA from a homozygous, wild-type β-globin individual (AA); from a heterozygous sickle β-globin individual (AS); and from a homozygous sickle β-globin individual (SS). Each DNA (50 ng genomic DNA to start each PCR) was analyzed in triplicate (3 pairs

of reactions each). The DNA type was reflected in the relative fluorescence intensities in each pair of completed amplifications. There was a significant increase in fluorescence only where a β-globin allele DNA matched the primer set. When measured on a spectrofluorometer (data not shown), this fluorescence was about three times that present in a PCR where both β-globin alleles were mismatched to the primer set. Gel electrophoresis (not shown) established that this increase in fluorescence was due to the synthesis of nearly a microgram of a DNA fragment of the expected size for β-globin. There was little synthesis of dsDNA in reactions in which the allele-specific primer was mismatched to both alleles.

Continuous monitoring of a PCR. Using a fiber optic device, it is possible to direct excitation illumination from a spectrofluorometer to a PCR undergoing thermocycling and to return its fluorescence to the spectrofluorometer. The fluorescence readout of such an arrangement, directed at an EtBr-containing amplification of Y-chromosome specific sequences from 25 ng of human male DNA, is shown in Figure 5. The readout from a control PCR with no target DNA is also shown. Thirty cycles of PCR were monitored for each.

The fluorescence trace as a function of time clearly shows the effect of the thermocycling. Fluorescence intensity rises and falls inversely with temperature. The fluorescence intensity is minimum at the denaturation temperature (94°C) and maximum at the annealing/extension temperature (50°C). In the negative-control PCR, these fluorescence maxima and minima do not change significantly over the thirty thermocycles, indicating that there is little dsDNA synthesis without the appropriate target DNA, and there is little if any bleaching of EtBr during the continuous illumination of the sample.

In the PCR containing male DNA, the fluorescence maxima at the annealing/extension temperature begin to increase at about 4000 seconds of thermocycling, and continue to increase with time, indicating that dsDNA is being produced at a detectable level. Note that the fluorescence minima at the denaturation temperature do not significantly increase, presumably because at this temperature there is no dsDNA for EtBr to bind. Thus the course of the amplification is followed by tracking the fluorescence increase at the annealing temperature. Analysis of the products of these two amplifications by gel electrophoresis showed a DNA fragment of the expected size for the male DNA containing sample and no detectable DNA synthesis for the control sample.

DISCUSSION

Downstream processes such as hybridization to a sequence-specific probe can enhance the specificity of DNA detection by PCR. The elimination of these processes means that the specificity of this homogeneous assay depends solely on that of PCR. In the case of sickle-cell disease, we have shown that PCR alone has sufficient DNA sequence specificity to permit genetic screening. Using appropriate amplification conditions, there is little non-specific production of dsDNA in the absence of the appropriate target allele.

The specificity required to detect pathogens can be more or less than that required to do genetic screening, depending on the number of pathogens in the sample and the amount of other DNA that must be taken with the sample. A difficult target is HIV, which requires detection of a viral genome that can be at the level of a few copies per thousands of host cells[6]. Compared with genetic screening, which is performed on cells containing at least one copy of the target sequence, HIV detection requires both more specificity and the input of more total

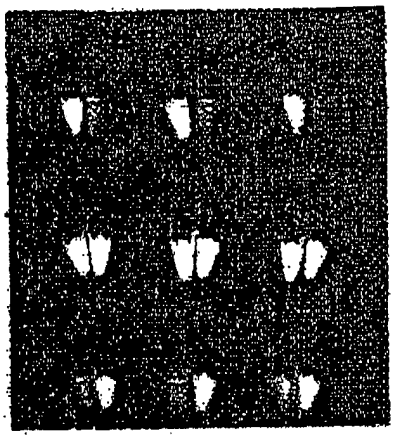


FIGURE 4 UV photography of PCR tubes containing amplifications using EtBr that are specific to wild-type (A) or sickle (S) alleles of the human β-globin gene. The left of each pair of tubes contains allele-specific primers to the wild-type alleles, the right tube primers to the sickle allele. The photograph was taken after 30 cycles of PCR, and the input DNAs and the alleles they contain are indicated. Fifty ng of DNA was used to begin PCR. Typing was done in triplicate (3 pairs of PCRs) for each input DNA.

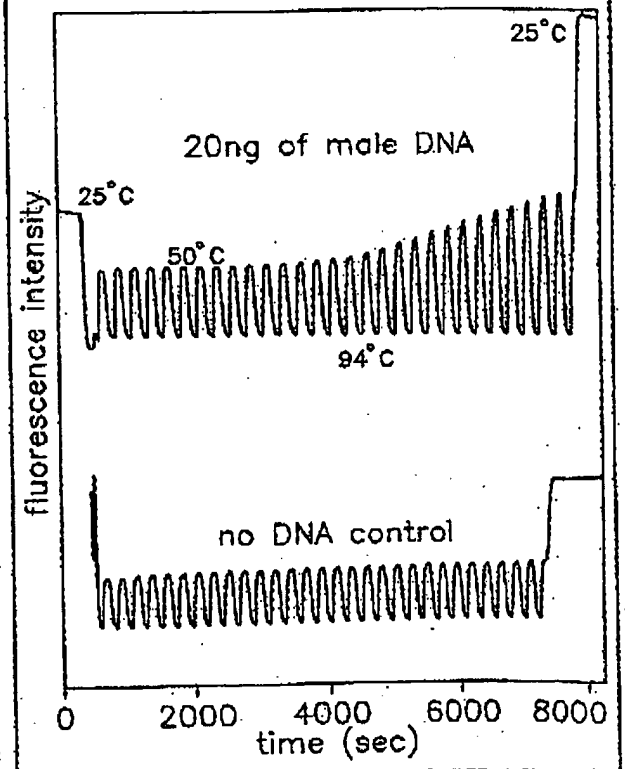


FIGURE 5 Continuous, real-time monitoring of a PCR. A fiber optic was used to carry excitation light to a PCR in progress and also emitted light back to a fluorometer (see Experimental Protocol). Amplification using human male-DNA specific primers in a PCR starting with 20 ng of human male DNA (top), or in a control PCR without DNA (bottom), were monitored. Thirty cycles of PCR were followed for each. The temperature cycled between 94°C (denaturation) and 50°C (annealing and extension). Note in the male DNA PCR, the cycle (time) dependent increase in fluorescence at the annealing/extension temperature.

DNA—up to microgram amounts—in order to have sufficient numbers of target sequences. This large amount of starting DNA in an amplification significantly increases the background fluorescence over which any additional fluorescence produced by PCR must be detected. An additional complication that occurs with targets in low copy-number is the formation of the "primer-dimer" artifact. This is the result of the extension of one primer using the other primer as a template. Although this occurs infrequently, once it occurs the extension product is a substrate for PCR amplification, and can compete with true PCR targets if those targets are rare. The primer-dimer product is of course dsDNA and thus is a potential source of false signal in this homogeneous assay.

To increase PCR specificity and reduce the effect of primer-dimer amplification, we are investigating a number of approaches, including the use of nested-primer amplifications that take place in a single tube[8], and the "hot-start", in which nonspecific amplification is reduced by raising the temperature of the reaction before DNA synthesis begins[23]. Preliminary results using these approaches suggest that primer-dimer is effectively reduced and it is possible to detect the increase in EtBr fluorescence in a PCR instigated by a single HIV genome in a background of 10^5 cells. With larger numbers of cells, the background fluorescence contributed by genomic DNA becomes problematic. To reduce this background, it may be possible to use sequence-specific DNA-binding dyes that can be made to preferentially bind PCR product over genomic DNA by incorporating the dye-binding DNA sequence into the PCR product through a 5' "add-on" to the oligonucleotide primer[24].

We have shown that the detection of fluorescence generated by an EtBr-containing PCR is straightforward, both once PCR is completed and continuously during thermocycling. The ease with which automation of specific DNA detection can be accomplished is the most promising aspect of this assay. The fluorescence analysis of completed PCRs is already possible with existing instrumentation in 96-well format[25]. In this format, the fluorescence in each PCR can be quantitated before, after, and even at selected points during thermocycling by moving the rack of PCRs to a 96-microwell plate fluorescence reader[26].

The instrumentation necessary to continuously monitor multiple PCRs simultaneously is also simple in principle. A direct extension of the apparatus used here is to have multiple fiberoptics transmit the excitation light and fluorescent emissions to and from multiple PCRs. The ability to monitor multiple PCRs continuously may allow quantitation of target DNA copy number. Figure 3 shows that the larger the amount of starting target DNA, the sooner during PCR a fluorescence increase is detected. Preliminary experiments (Higuchi and Dollinger, manuscript in preparation) with continuous monitoring have shown a sensitivity to two-fold differences in initial target DNA concentration.

Conversely, if the number of target molecules is known—as it can be in genetic screening—continuous monitoring may provide a means of detecting false positive and false negative results. With a known number of target molecules, a true positive would exhibit detectable fluorescence by a predictable number of cycles of PCR. Increases in fluorescence detected before or after that cycle would indicate potential artifacts. False negative results due to, for example, inhibition of DNA polymerase, may be detected by including within each PCR an inefficiently amplifying marker. This marker results in a fluorescence increase only after a large number of cycles—many more than are necessary to detect a true positive. If a sample fails to have a fluorescence increase after this many cycles, inhibition may be suspected. Since, in this assay, conclusions are drawn based on the presence or absence of fluorescence signal alone, such controls may be important. In any event, before any test based on this principle is ready for the clinic, an assessment of its false positive/false negative rates will need to be obtained using a large number of known samples.

In summary, the inclusion in PCR of dyes whose fluorescence is enhanced upon binding dsDNA makes it possible to detect specific DNA amplification from outside the PCR tube. In the future, instruments based upon this principle may facilitate the more widespread use of PCR in applications that demand the high throughput of samples.

EXPERIMENTAL PROTOCOL

Human HLA-DQα gene amplifications containing EtBr. PCRs were set up in 100 µl volumes containing 10 mM Tris-HCl, pH 8.3; 50 mM KCl; 4 mM $MgCl_2$; 2.5 units of *Taq* DNA polymerase (Perkin-Elmer Cetus, Norwalk, CT); 20 pmole each of human HLA-DQα gene specific oligonucleotide primers (GH26 and GH27[19] and approximately 10^7 copies of DQα PCR product diluted from a previous reaction. Ethidium bromide (EtBr; Sigma) was used at the concentrations indicated in Figure 2. Thermocycling proceeded for 20 cycles in a model 480 thermocycler (Perkin-Elmer Cetus, Norwalk, CT) using a "step-cycle" program of 94°C for 1 min. denaturation and 60°C for 30 sec. annealing and 72°C for 30 sec. extension.

Y-chromosome specific PCR. PCRs (100 µl total reaction volume) containing 0.5 µg/ml EtBr were prepared as described for HLA-DQα, except with different primers and target DNAs. These PCRs contained 15 pmole each male DNA-specific primers Y1.1 and Y1.2[20], and either 60 ng male, 60 ng female, 2 ng male, or no human DNA. Thermocycling was 94°C for 1 min. and 60°C for 1 min using a "step-cycle" program. The number of cycles for a sample were as indicated in Figure 3. Fluorescence measurement is described below.

Allele-specific, human β-globin gene PCR. Amplifications of 100 µl volume using 0.5 µg/ml of EtBr were prepared as described for HLA-DQα above except with different primers and target DNAs. These PCRs contained either primer pair HGP2/Hβ14A (wild-type globin specific primers) or HGP2/Hβ14S (sickle-globin specific primers) at 10 pmole each primer per PCR. These primers were developed by Wu et al.[21]. Three different target DNAs were used in separate amplifications—50 ng each of human DNA that was homozygous for the sickle trait (SS), DNA that was heterozygous for the sickle trait (AS), or DNA that was homozygous for the w.t. globin (AA). Thermocycling was for 30 cycles at 94°C for 1 min. and 55°C for 1 min. using a "step-cycle" program. An annealing temperature of 55°C had been shown by Wu et al.[21] to provide allele-specific amplification. Completed PCRs were photographed through a red filter (Wratten 23A) after placing the reaction tubes atop a model TM-36 transilluminator (UV-products San Gabriel, CA).

Fluorescence measurement. Fluorescence measurements were made on PCRs containing EtBr in a Fluorolog-2 fluorometer (SPEX, Edison, NJ). Excitation was at the 500 nm band with about 2 nm bandwidth with a GG 435 nm cut-off filter (Melles Grist, Inc., Irvine, CA) to exclude second-order light. Emitted light was detected at 570 nm with a bandwidth of about 7 nm. An OG 530 nm cut-off filter was used to remove the excitation light.

Continuous fluorescence monitoring of PCR. Continuous monitoring of a PCR in progress was accomplished using the spectrofluorometer and settings described above as well as a fiberoptic accessory (SPEX cat. no. 1950) to both send excitation light to, and receive emitted light from, a PCR placed in a well of a model 480 thermocycler (Perkin-Elmer Cetus). The probe end of the fiberoptic cable was attached with "5 minute-epoxy" to the open top of a PCR tube (a 0.5 ml polypropylene centrifuge tube with its cap removed) effectively sealing it. The exposed top of the PCR tube and the end of the fiberoptic cable were shielded from room light and the room lights were kept dimmed during each run. The monitored PCR was an amplification of Y-chromosome-specific repeat sequences as described above, except using an annealing/extension temperature of 50°C. The reaction was covered with mineral oil (2 drops) to prevent evaporation. Thermocycling and fluorescence measurement were started simultaneously. A time-base scan with a 10 second integration time

was used and the emission signal was ratioed to the excitation signal to control for changes in light-source intensity. Data were collected using the drm3000f, version 2.5 (SPEX) data system.

Acknowledgments

We thank Bob Jones for help with the spectrofluormetric measurements and Heatherbell Fong for editing this manuscript.

References

1. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. 1986. Specific enzymatic amplification of DNA *in vitro*: The polymerase chain reaction. CSHSQB 51:263–273.
2. White, T. J., Arnheim, N. and Erlich, H. A. 1989. The polymerase chain reaction. Trends Genet. 5:185–189.
3. Erlich, H. A., Gelfand, D. and Sninsky, J. J. 1991. Recent advances in the polymerase chain reaction. Science 252:1643–1651.
4. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491.
5. Saiki, R. K., Walsh, P. S., Levenson, C. H. and Erlich, H. A. 1989. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc. Natl. Acad. Sci. USA 86:6230–6234.
6. Kwok, S. Y., Mack, D. H., Mullis, K. B., Poiesz, B. J., Ehrlich, G. D., Blair, D. and Friedman-Kien, A. S. 1987. Identification of human immunodeficiency virus sequences by using *in vitro* enzymatic amplification and oligomer cleavage detection. J. Virol. 61:1690–1694.
7. Chehab, F. F., Doherty, M., Cai, S. P., Kan, Y. W., Cooper, S. and Rubin, E. M. 1987. Detection of sickle cell anemia and thalassemias. Nature 329:293–294.
8. Horn, G. T., Richards, B. and Klinger, K. W. 1989. Amplification of a highly polymorphic VNTR segment by the polymerase chain reaction. Nuc. Acids Res. 16:2140.
9. Katz, E. D. and Dong, M. W. 1990. Rapid analysis and purification of polymerase chain reaction products by high-performance liquid chromatography. Biotechniques 8:546–555.
10. Heiger, D. N., Cohen, A. S. and Karger, B. L. 1990. Separation of DNA restriction fragments by high performance capillary electrophoresis with low and zero crosslinked polyacrylamide using continuous and pulsed electric fields. J. Chromatogr. 516:33–48.
11. Kwok, S. Y. and Higuchi, R. G. 1989. Avoiding false positives with PCR. Nature 339:237–238.
12. Chehab, F. F. and Kan, Y. W. 1989. Detection of specific DNA sequences by fluorescence amplification: a color complementation assay. Proc. Natl. Acad. Sci. USA 86:9178–9182.
13. Holland, P. M., Abramson, R. D., Watson, R. and Gelfand, D. H. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc. Natl. Acad. Sci. USA 88:7276–7280.
14. Markovits, J., Roques, B. P. and Le Pecq, J. B. 1979. Ethidium dimer: a new reagent for the fluorimetric determination of nucleic acids. Anal. Biochem. 94:259–264.
15. Kapuscinski, J. and Szer, W. 1979. Interactions of 4',6-diamidino-2-phenylindole with synthetic polynucleotides. Nuc. Acids Res. 6:3519–3534.
16. Searle, M. S. and Embrey, K. J. 1990. Sequence-specific interaction of Hoescht 33258 with the minor groove of an adenine-tract DNA duplex studied in solution by ^{1}H NMR spectroscopy. Nuc. Acids Res. 18:3753–3762.
17. Li, H. H., Gyllensten, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A. and Arnheim, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature 335:414–417.
18. Abbott, M. A., Poiesz, B. J., Byrne, B. C., Kwok, S. Y., Sninsky, J. J. and Erlich, H. A. 1988. Enzymatic gene amplification: qualitative and quantitative methods for detecting proviral DNA amplified *in vitro*. J. Infect. Dis. 158:1158.
19. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. and Erlich, H. A. 1986. Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. Nature 324:163–166.
20. Kogan, S. C., Doherty, M. and Gitschier, J. 1987. An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. N. Engl. J. Med. 317:985–990.
21. Wu, D. Y., Ugozzoli, L., Pal, B. K. and Wallace, R. B. 1989. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. Proc. Natl. Acad. Sci. USA 86:2757–2760.
22. Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. Nuc. Acids Res. 18:999–1005.
23. Chou, Q., Russell, M., Birch, D., Raymond, J. and Bloch, W. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Submitted.
24. Higuchi, R. 1989. Using PCR to engineer DNA. p. 61–70. *In:* PCR Technology. H. A. Erlich (Ed.). Stockton Press, New York, N.Y.
25. Hatt, L., Atwood, J. G., DiCesare, J., Katz, E., Pizzara, E., Williams, J. F. and Woudenberg, T. 1991. A high-performance system for automation of the polymerase chain reaction. Biotechniques 10:102–108, 106–112.
26. Tumosa, N. and Kahan, L. 1989. Fluorescent EIA screening of monoclonal antibodies to cell surface antigens. J. Immun. Meth. 116:59–63.

Oligonucleotides with Fluorescent Dyes at Opposite Ends Provide a Quenched Probe System Useful for Detecting PCR Product and Nucleic Acid Hybridization

Kenneth J. Livak, Susan J.A. Flood, Jeffrey Marmaro, William Giusti, and Karin Deetz

Perkin-Elmer, Applied Biosystems Division, Foster City, California 94404

The 5′ nuclease PCR assay detects the accumulation of specific PCR product by hybridization and cleavage of a double-labeled fluorogenic probe during the amplification reaction. The probe is an oligonucleotide with both a reporter fluorescent dye and a quencher dye attached. An increase in reporter fluorescence intensity indicates that the probe has hybridized to the target PCR product and has been cleaved by the 5′→3′ nucleolytic activity of *Taq* DNA polymerase. In this study, probes with the quencher dye attached to an internal nucleotide were compared with probes with the quencher dye attached to the 3′-end nucleotide. In all cases, the reporter dye was attached to the 5′ end. All intact probes showed quenching of the reporter fluorescence. In general, probes with the quencher dye attached to the 3′-end nucleotide exhibited a larger signal in the 5′ nuclease PCR assay than the internally labeled probes. It is proposed that the larger signal is caused by increased likelihood of cleavage by *Taq* DNA polymerase when the probe is hybridized to a template strand during PCR. Probes with the quencher dye attached to the 3′-end nucleotide also exhibited an increase in reporter fluorescence intensity when hybridized to a complementary strand. Thus, oligonucleotides with reporter and quencher dyes attached at opposite ends can be used as homogeneous hybridization probes.

A homogeneous assay for detecting the accumulation of specific PCR product that uses a double-labeled fluorogenic probe was described by Lee et al.[1] The assay exploits the 5′→3′ nucleolytic activity of *Taq* DNA polymerase[2,3] and is diagramed in Figure 1. The fluorogenic probe consists of an oligonucleotide with a reporter fluorescent dye, such as a fluorescein, attached to the 5′ end; and a quencher dye, such as a rhodamine, attached internally. When the fluorescein is excited by irradiation, its fluorescent emission will be quenched if the rhodamine is close enough to be excited through the process of fluorescence energy transfer (FET).[4,5] During PCR, if the probe is hybridized to a template strand, *Taq* DNA polymerase will cleave the probe because of its inherent 5′→3′ nucleolytic activity. If the cleavage occurs between the fluorescein and rhodamine dyes, it causes an increase in fluorescein fluorescence intensity because the fluorescein is no longer quenched. The increase in fluorescein fluorescence intensity indicates that the probe-specific PCR product has been generated. Thus, FET between a reporter dye and a quencher dye is critical to the performance of the probe in the 5′ nuclease PCR assay.

Quenching is completely dependent on the physical proximity of the two dyes.[6] Because of this, it has been assumed that the quencher dye must be attached near the 5′ end. Surprisingly, we have found that attaching a rhodamine dye at the 3′ end of a probe still provides adequate quenching for the probe to perform in the 5′ nuclease PCR assay. Furthermore, cleavage of this type of probe is not required to achieve some reduction in quenching. Oligonucleotides with a reporter dye on the 5′ end and a quencher dye on the 3′ end exhibit a much higher reporter fluorescence when double-stranded as compared with single-stranded. This should make it possible to use this type of double-labeled probe for homogeneous detection of nucleic acid hybridization.

MATERIALS AND METHODS

Oligonucleotides

Table 1 shows the nucleotide sequence of the oligonucleotides used in this study. Linker arm nucleotide (LAN) phosphoramidite was obtained from Glen Research. The standard DNA phosphoramidites, 6-carboxyfluorescein (6-FAM) phosphoramidite, 6-carboxytetramethylrhodamine succinimidyl ester (TAMRA NHS ester), and Phosphalink for attaching a 3′-blocking phosphate, were obtained from Perkin-Elmer, Applied Biosystems Division. Oligonucleotide synthesis was performed using an ABI model 394 DNA synthesizer (Applied Biosystems). Primer and complement oligonucleotides were purified using Oligo Purification Cartridges (Applied Biosystems). Double-labeled probes were synthesized with 6-FAM-labeled phosphoramidite at the 5′ end, LAN replacing one of the T's in the sequence, and Phosphalink at the 3′ end. Following deprotection and ethanol precipitation, TAMRA NHS ester was coupled to the LAN-containing oligonucleotide in 250

4:357–362©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

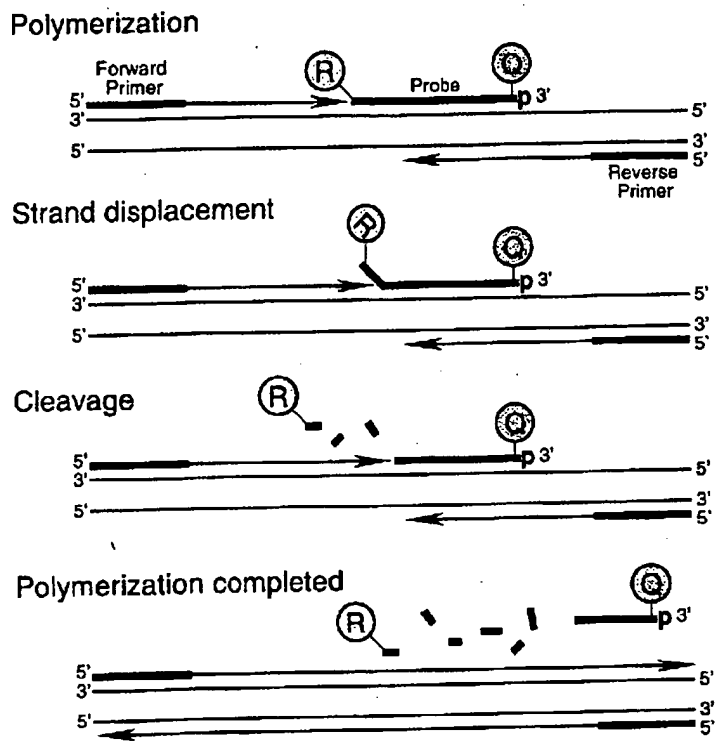


FIGURE 1 Diagram of 5' nuclease assay. Stepwise representation of the 5' → 3' nucleolytic activity of *Taq* DNA polymerase acting on a fluorogenic probe during one extension phase of PCR.

mM Na-bicarbonate buffer (pH 9.0) at room temperature. Unreacted dye was removed by passage over a PD-10 Sephadex column. Finally, the double-labeled probe was purified by preparative high-performance liquid chromatography (HPLC) using an Aquapore C_8 220×4.6-mm column with 7-μm particle size. The column was developed with a 24-min linear gradient of 8–20% acetonitrile in 0.1 M TEAA (triethylamine acetate). Probes are named by designating the sequence from Table 1 and the position of the LAN–TAMRA moiety. For example, probe A1-7 has sequence A1 with LAN–TAMRA at nucleotide position 7 from the 5' end.

PCR Systems

All PCR amplifications were performed in the Perkin-Elmer GeneAmp PCR System 9600 using 50-μl reactions that contained 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 400 μM dUTP, 0.5 unit of AmpErase uracil *N*-glycosylase (Perkin-Elmer), and 1.25 unit of AmpliTaq DNA polymerase (Perkin-Elmer). A 295-bp segment from exon 3 of the human β-actin gene (nucleotides 2141–2435 in the sequence of Nakajima-Iijima et al.)[7] was amplified using primers AFP and ARP (Table 1), which are modified slightly from those of du Breuil et al.[8] Actin amplification reactions contained 4 mM $MgCl_2$, 20 ng of human genomic DNA, 50 nM A1 or A3 probe, and 300 nM each primer. The thermal regimen was 50°C (2 min), 95°C (10 min), 40 cycles of 95°C (20 sec), 60°C (1 min), and hold at 72°C. A 515-bp segment was amplified from a plasmid that consists of a segment of λ DNA (nucleotides 32,220–32,747) inserted in the *Sma*I site of vector pUC119. These reactions contained 3.5 mM $MgCl_2$, 1 ng of plasmid DNA, 50 nM P2 or P5 probe, 200 nM primer F119, and 200 nM primer R119. The thermal regimen was 50°C (2 min), 95°C (10 min), 25 cycles of 95°C (20 sec), 57°C (1 min), and hold at 72°C.

Fluorescence Detection

For each amplification reaction, a 40-μl aliquot of a sample was transferred to an individual well of a white, 96-well microtiter plate (Perkin-Elmer). Fluorescence was measured on the Perkin-Elmer TaqMan LS-50B System, which consists of a luminescence spectrometer with plate reader assembly, a 485-nm excitation filter, and a 515-nm emission filter. Excitation was at 488 nm using a 5-nm slit width. Emission was measured at 518 nm for 6-FAM (the reporter or R value) and 582 nm for TAMRA (the quencher or Q value) using a 10-nm slit width. To determine the increase in reporter emission that is caused by cleavage of the probe during PCR, three normalizations are applied to the raw emission data. First, emission intensity of a buffer blank is subtracted for each wavelength. Second, emission intensity of the reporter is

TABLE 1 Sequences of Oligonucleotides

Name	Type	Sequence
F119	primer	ACCCACAGGAACTGATCACCACTC
R119	primer	ATGTCGCGTTCCGGCTGACGTTCTGC
P2	probe	TCGCATTACTGATCGTTGCCAACCAGTp
P2C	complement	GTACTGGTTGGCAACGATCAGTAATGCGATG
P5	probe	CGGATTTGCTGGTATCTATGACAAGGATp
P5C	complement	TTCATCCTTGTCATAGATACCAGCAAATCCG
AFP	primer	TCACCCACACTGTGCCCATCTACGA
ARP	primer	CAGCGGAACCGCTCATTGCCAATGG
A1	probe	ATGCCCTCCCCCATGCCATCCTGCGTp
A1C	complement	AGACGCAGGATGGCATGGGGGAGGGCATAC
A3	probe	CGCCCTGGACTTCGAGCAAGAGATp
A3C	complement	CCATCTCTTGCTCGAAGTCCAGGGCGAC

For each oligonucleotide used in this study, the nucleic acid sequence is given, written in the 5' → 3' direction. There are three types of oligonucleotides: PCR primer, fluorogenic probe used in the 5' nuclease assay, and complement used to hybridize to the corresponding probe. For the probes, the underlined base indicates a position where LAN with TAMRA attached was substituted for a T. (p) The presence of a 3' phosphate on each probe.

A1-2	RAQGCCCTCCCCCATGCCATCCTGCGTp
A1-7	RATGCCCQCCCCCATGCCATCCTGCGTp
A1-14	RATGCCCTCCCCCAQGCCATCCTGCGTp
A1-19	RATGCCCTCCCCCATGCCAQCCTGCGTp
A1-22	RATGCCCTCCCCCATGCCATCCQGCGTp
A1-26	RATGCCCTCCCCCATGCCATCCTGCGQp

Probe	518 nm		582 nm		RQ⁻	RQ⁺	ΔRQ
	no temp.	+ temp.	no temp.	+ temp.			
A1-2	25.5 ± 2.1	32.7 ± 1.9	38.2 ± 3.0	38.2 ± 2.0	0.67 ± 0.01	0.86 ± 0.06	0.19 ± 0.06
A1-7	53.5 ± 4.3	395.1 ± 21.4	108.5 ± 6.3	110.3 ± 5.3	0.49 ± 0.03	3.58 ± 0.17	3.09 ± 0.18
A1-14	127.0 ± 4.9	403.5 ± 19.1	109.7 ± 5.3	93.1 ± 6.3	1.16 ± 0.02	4.34 ± 0.15	3.18 ± 0.15
A1-19	187.5 ± 17.9	422.7 ± 7.7	70.3 ± 7.4	73.0 ± 2.8	2.67 ± 0.05	5.80 ± 0.15	3.13 ± 0.16
A1-22	224.6 ± 9.4	482.2 ± 43.6	100.0 ± 4.0	96.2 ± 9.6	2.25 ± 0.03	5.02 ± 0.11	2.77 ± 0.12
A1-26	160.2 ± 8.9	454.1 ± 18.4	93.1 ± 5.4	90.7 ± 3.2	1.72 ± 0.02	5.01 ± 0.08	3.29 ± 0.08

FIGURE 2 Results of 5′ nuclease assay comparing β-actin probes with TAMRA at different nucleotide positions. As described in Materials and Methods, PCR amplifications containing the indicated probes were performed, and the fluorescence emission was measured at 518 and 582 nm. Reported values are the average ± 1 S.D. for six reactions run without added template (no temp.) and six reactions run with template (+temp.). The RQ ratio was calculated for each individual reaction and averaged to give the reported RQ^- and RQ^+ values.

divided by the emission intensity of the quencher to give an RQ ratio for each reaction tube. This normalizes for well-to-well variations in probe concentration and fluorescence measurement. Finally, ΔRQ is calculated by subtracting the RQ value of the no-template control (RQ^-) from the RQ value for the complete reaction including template (RQ^+).

RESULTS

A series of probes with increasing distances between the fluorescein reporter and rhodamine quencher were tested to investigate the minimum and maximum spacing that would give an acceptable performance in the 5′ nuclease PCR assay. These probes hybridize to a target sequence in the human β-actin gene. Figure 2 shows the results of an experiment in which these probes were included in PCR that amplified a segment of the β-actin gene containing the target sequence. Performance in the 5′ nuclease PCR assay is monitored by the magnitude of ΔRQ, which is a measure of the increase in reporter fluorescence caused by PCR amplification of the probe target. Probe A1-2 has a ΔRQ value that is close to zero, indicating that the probe was not cleaved appreciably during the amplification reaction. This suggests that with the quencher dye on the second nucleotide from the 5′ end, there is insufficient room for *Taq* polymerase to cleave efficiently between the reporter and quencher. The other five probes exhibited comparable ΔRQ values that are clearly different from zero. Thus, all five probes are being cleaved during PCR amplification resulting in a similar increase in reporter fluorescence. It should be noted that complete digestion of a probe produces a much larger increase in reporter fluorescence than that observed in Figure 2 (data not shown). Thus, even in reactions where amplification occurs, the majority of probe molecules remain uncleaved. It is mainly for this reason that the fluorescence intensity of the quencher dye TAMRA changes little with amplification of the target. This is what allows us to use the 582-nm fluorescence reading as a normalization factor.

The magnitude of RQ^- depends mainly on the quenching efficiency inherent in the specific structure of the probe and the purity of the oligonucleotide. Thus, the larger RQ^- values indicate that probes A1-14, A1-19, A1-22, and A1-26 probably have reduced quenching as compared with A1-7. Still, the degree of quenching is sufficient to detect a highly significant increase in reporter fluorescence when each of these probes is cleaved during PCR.

To further investigate the ability of TAMRA on the 3′ end to quench 6-FAM on the 5′ end, three additional pairs of probes were tested in the 5′ nuclease PCR assay. For each pair, one probe has TAMRA attached to an internal nucleotide and the other has TAMRA attached to the 3′ end nucleotide. The results are shown in Table 2. For all three sets, the probe with the 3′ quencher exhibits a ΔRQ value that is considerably higher than for the probe with the internal quencher. The RQ^- values suggest that differences in quenching are not as great as those observed with some of the A1 probes. These results demonstrate that a quencher dye on the 3′ end of an oligonucleotide can quench efficiently the

TABLE 2 Results of 5′ Nuclease Assay Comparing Probes with TAMRA Attached to an Internal or 3′-terminal Nucleotide

Probe	518 nm		582 nm		RQ⁻	RQ⁺	ΔRQ
	no temp.	+ temp.	no temp.	+ temp.			
A3-6	54.6 ± 3.2	84.8 ± 3.7	116.2 ± 6.4	115.6 ± 2.5	0.47 ± 0.02	0.73 ± 0.03	0.26 ± 0.04
A3-24	72.1 ± 2.9	236.5 ± 11.1	84.2 ± 4.0	90.2 ± 3.8	0.86 ± 0.02	2.62 ± 0.05	1.76 ± 0.05
P2-7	82.8 ± 4.4	384.0 ± 34.1	105.1 ± 6.4	120.4 ± 10.2	0.79 ± 0.02	3.19 ± 0.16	2.40 ± 0.16
P2-27	113.4 ± 6.6	555.4 ± 14.1	140.7 ± 8.5	118.7 ± 4.8	0.81 ± 0.01	4.68 ± 0.10	3.88 ± 0.10
P5-10	77.5 ± 6.5	244.4 ± 15.9	86.7 ± 4.3	95.8 ± 6.7	0.89 ± 0.05	2.55 ± 0.06	1.66 ± 0.08
P5-28	64.0 ± 5.2	333.6 ± 12.1	100.6 ± 6.1	94.7 ± 6.3	0.63 ± 0.02	3.53 ± 0.12	2.89 ± 0.13

Reactions containing the indicated probes and calculations were performed as described in Material and Methods and in the legend to Fig. 2.

fluorescence of a reporter dye on the 5′ end. The degree of quenching is sufficient for this type of oligonucleotide to be used as a probe in the 5′ nuclease PCR assay.

To test the hypothesis that quenching by a 3′ TAMRA depends on the flexibility of the oligonucleotide, fluorescence was measured for probes in the single-stranded and double-stranded states. Table 3 reports the fluorescence observed at 518 and 582 nm. The relative degree of quenching is assessed by calculating the RQ ratio. For probes with TAMRA 6–10 nucleotides from the 5′ end, there is little difference in the RQ values when comparing single-stranded with double-stranded oligonucleotides. The results for probes with TAMRA at the 3′ end are much different. For these probes, hybridization to a complementary strand causes a dramatic increase in RQ. We propose that this loss of quenching is caused by the rigid structure of double-stranded DNA, which prevents the 5′ and 3′ ends from being in proximity.

When TAMRA is placed toward the 3′ end, there is a marked Mg^{2+} effect on quenching. Figure 3 shows a plot of observed RQ values for the A1 series of probes as a function of Mg^{2+} concentration. With TAMRA attached near the 5′ end (probe A1-2 or A1-7), the RQ value at 0 mM Mg^{2+} is only slightly higher than RQ at 10 mM Mg^{2+}. For probes A1-19, A1-22, and A1-26, the RQ values at 0 mM Mg^{2+} are very high, indicating a much reduced quenching efficiency. For each of these probes, there is a marked decrease in RQ at 1 mM Mg^{2+} followed by a gradual decline as the Mg^{2+} concentration increases to 10 mM. Probe A1-14 shows an intermediate RQ value at 0 mM Mg^{2+} with a gradual decline at higher Mg^{2+} concentrations. In a low-salt environment with no Mg^{2+} present, a single-stranded oligonucleotide would be expected to adopt an extended conformation because of electrostatic repulsion. The binding of Mg^{2+} ions acts to shield the negative charge of the phosphate backbone so that the oligonucleotide can adopt conformations where the 3′ end is close to the 5′ end. Therefore, the observed Mg^{2+} effects support the notion that quenching of a 5′ reporter dye by TAMRA at or near the 3′ end depends on the flexibility of the oligonucleotide.

DISCUSSION

The striking finding of this study is that it seems the rhodamine dye TAMRA, placed at any position in an oligonucleotide, can quench the fluorescent emission of a fluorescein (6-FAM) placed at the 5′ end. This implies that a single-stranded, double-labeled oligonucleotide must be able to adopt conformations where the TAMRA is close to the 5′ end. It should be noted that the decay of 6-FAM in the excited state requires a certain amount of time. Therefore, what matters for quenching is not the average distance between 6-FAM and TAMRA but, rather, how close TAMRA can get to 6-FAM during the lifetime of the 6-FAM excited state. As long as the decay time of the excited state is relatively long compared with the molecular motions of the oligonucleotide, quenching can occur. Thus, we propose that TAMRA at the 3′ end, or any other position, can quench 6-FAM at the 5′ end because TAMRA is in proximity to 6-FAM often enough to be able to accept energy transfer from an excited 6-FAM.

Details of the fluorescence measurements remain puzzling. For example, Table 3 shows that hybridization of probes A1-26, A3-24, and P5-28 to their complementary strands not only causes a large increase in 6-FAM fluorescence at 518 nm but also causes a modest increase in TAMRA fluorescence at 582 nm. If TAMRA is being excited by energy transfer from quenched 6-FAM, then loss of quenching attributable to hybridization should cause a decrease in the fluorescence emission of TAMRA. The fact that the fluorescence emission of TAMRA increases indicates that the situation is more complex. For example, we have anecdotal evidence that the bases of the oligonucleotide, especially G, quench the fluorescence of both 6-FAM and TAMRA to some degree. When double-stranded, base-pairing may reduce the ability of the bases to quench. The primary factor causing the quenching of 6-FAM in an intact probe is the TAMRA dye. Evidence for the importance of TAMRA is that 6-FAM fluorescence remains relatively unchanged when probes labeled only with 6-FAM are used in the 5′ nuclease PCR assay (data not shown). Secondary effectors of fluorescence, both before and after cleavage of the probe, need to be explored further.

Regardless of the physical mechanism, the relative independence of position and quenching greatly simplifies the design of probes for the 5′ nuclease PCR assay. There are three main factors that determine the performance of a double-labeled fluorescent probe in the 5′ nuclease PCR assay. The first factor is the degree of quenching observed in the intact probe. This is characterized by the value of RQ^-, which is the ratio of reporter to quencher fluorescent emissions for a no template control PCR. Influences on the value of RQ^- include the particular reporter and quencher

TABLE 3 Comparison of Fluorescence Emissions of Single-stranded and Double-stranded Fluorogenic Probes

	518 nm		582 nm		RQ	
Probe	ss	ds	ss	ds	ss	ds
A1-7	27.75	68.53	61.08	138.18	0.45	0.50
A1-26	43.31	509.38	53.50	93.86	0.81	5.43
A3-6	16.75	62.88	39.33	165.57	0.43	0.38
A3-24	30.05	578.64	67.72	140.25	0.45	3.21
P2-7	35.02	70.13	54.63	121.09	0.64	0.58
P2-27	39.89	320.47	65.10	61.13	0.61	5.25
P5-10	27.34	144.85	61.95	165.54	0.44	0.87
P5-28	33.65	462.29	72.39	104.61	0.46	4.43

(ss) Single-stranded. The fluorescence emissions at 518 or 582 nm for solutions containing a final concentration of 50 nM indicated probe, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, and 10 mM $MgCl_2$. (ds) Double-stranded. The solutions contained, in addition, 100 nM A1C for probes A1-7 and A1-26, 100 nM A3C for probes A3-6 and A3-24, 100 nM P2C for probes P2-7 and P2-27, or 100 nM P5C for probes P5-10 and P5-28. Before the addition of $MgCl_2$, 120 µl of each sample was heated at 95°C for 5 min. Following the addition of 80 µl of 25 mM $MgCl_2$, each sample was allowed to cool to room temperature and the fluorescence emissions were measured. Reported values are the average of three determinations.

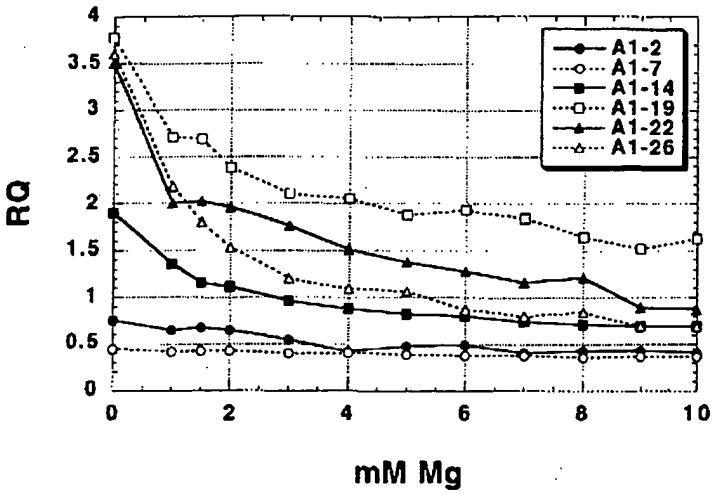


FIGURE 3 Effect of Mg^{2+} concentration on RQ ratio for the A1 series of probes. The fluorescence emission intensity at 518 and 582 nm was measured for solutions containing 50 nM probe, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, and varying amounts (0–10 mM) of $MgCl_2$. The calculated RQ ratios (518 nm intensity divided by 582 nm intensity) are plotted vs. $MgCl_2$ concentration (mM Mg). The key (*upper right*) shows the probes examined.

dyes used, spacing between reporter and quencher dyes, nucleotide sequence context effects, presence of structure or other factors that reduce flexibility of the oligonucleotide, and purity of the probe. The second factor is the efficiency of hybridization, which depends on probe T_m, presence of secondary structure in probe or template, annealing temperature, and other reaction conditions. The third factor is the efficiency at which *Taq* DNA polymerase cleaves the bound probe between the reporter and quencher dyes. This cleavage is dependent on sequence complementarity between probe and template as shown by the observation that mismatches in the segment between reporter and quencher dyes drastically reduce the cleavage of probe.[1]

The rise in RQ^- values for the A1 series of probes seems to indicate that the degree of quenching is reduced somewhat as the quencher is placed toward the 3′ end. The lowest apparent quenching is observed for probe A1-19 (see Fig. 3) rather than for the probe where the TAMRA is at the 3′ end (A1-26). This is understandable, as the conformation of the 3′ end position would be expected to be less restricted than the conformation of an internal position. In effect, a quencher at the 3′ end is freer to adopt conformations close to the 5′ reporter dye than is an internally placed quencher. For the other three sets of probes, the interpretation of RQ^- values is less clear-cut. The A3 probes show the same trend as A1, with the 3′ TAMRA probe having a larger RQ^- than the internal TAMRA probe. For the P2 pair, both probes have about the same RQ^- value. For the P5 probes, the RQ^- for the 3′ probe is less than for the internally labeled probe. Another factor that may explain some of the observed variation is that purity affects the RQ^- value. Although all probes are HPLC purified, a small amount of contamination with unquenched reporter can have a large effect on RQ^-.

Although there may be a modest effect on degree of quenching, the position of the quencher apparently can have a large effect on the efficiency of probe cleavage. The most drastic effect is observed with probe A1-2, where placement of the TAMRA on the second nucleotide reduces the efficiency of cleavage to almost zero. For the A3, P2, and P5 probes, ΔRQ is much greater for the 3′ TAMRA probes as compared with the internal TAMRA probes. This is explained most easily by assuming that probes with TAMRA at the 3′ end are more likely to be cleaved between reporter and quencher than are probes with TAMRA attached internally. For the A1 probes, the cleavage efficiency of probe A1-7 must already be quite high, as ΔRQ does not increase when the quencher is placed closer to the 3′ end. This illustrates the importance of being able to use probes with a quencher on the 3′ end in the 5′ nuclease PCR assay. In this assay, an increase in the intensity of reporter fluorescence is observed only when the probe is cleaved between the reporter and quencher dyes. By placing the reporter and quencher dyes on the opposite ends of an oligonucleotide probe, any cleavage that occurs will be detected. When the quencher is attached to an internal nucleotide, sometimes the probe works well (A1-7) and other times not so well (A3-6). The relatively poor performance of probe A3-6 presumably means the probe is being cleaved 3′ to the quencher rather than between the reporter and quencher. Therefore, the best chance of having a probe that reliably detects accumulation of PCR product in the 5′ nuclease PCR assay is to use a probe with the reporter and quencher dyes on opposite ends.

Placing the quencher dye on the 3′ end may also provide a slight benefit in terms of hybridization efficiency. The presence of a quencher attached to an internal nucleotide might be expected to disrupt base-pairing and reduce the T_m of a probe. In fact, a 2°C–3°C reduction in T_m has been observed for two probes with internally attached TAMRAs.[9] This disruptive effect would be minimized by placing the quencher at the 3′ end. Thus, probes with 3′ quenchers might exhibit slightly higher hybridization efficiencies than probes with internal quenchers.

The combination of increased cleavage and hybridization efficiencies means that probes with 3′ quenchers probably will be more tolerant of mismatches between probe and target as compared with internally labeled probes. This tolerance of mismatches can be advantageous, as when trying to use a single probe to detect PCR-amplified products from samples of different species. Also, it means that cleavage of probe during PCR is less sensitive to alterations in annealing temperature or other reaction conditions. The one application where tolerance of mismatches may be a disadvantage is for allelic discrimination. Lee et al.[1] demonstrated that allele-specific probes were cleaved between reporter and quencher only when hybridized to a perfectly complementary target. This allowed them to distinguish the normal human cystic fibrosis allele from the *ΔF508* mutant. Their probes had TAMRA attached to the seventh nucleotide from

the 5' end and were designed so that any mismatches were between the reporter and quencher. Increasing the distance between reporter and quencher would lessen the disruptive effect of mismatches and allow cleavage of the probe on the incorrect target. Thus, probes with a quencher attached to an internal nucleotide may still be useful for allelic discrimination.

In this study loss of quenching upon hybridization was used to show that quenching by a 3' TAMRA is dependent on the flexibility of a single-stranded oligonucleotide. The increase in reporter fluorescence intensity, though, could also be used to determine whether hybridization has occurred or not. Thus, oligonucleotides with reporter and quencher dyes attached at opposite ends should also be useful as hybridization probes. The ability to detect hybridization in real time means that these probes could be used to measure hybridization kinetics. Also, this type of probe could be used to develop homogeneous hybridization assays for diagnostics or other applications. Bagwell et al.[10] describe just this type of homogeneous assay where hybridization of a probe causes an increase in fluorescence caused by a loss of quenching. However, they utilized a complex probe design that requires adding nucleotides to both ends of the probe sequence to form two imperfect hairpins. The results presented here demonstrate that the simple addition of a reporter dye to one end of an oligonucleotide and a quencher dye to the other end generates a fluorogenic probe that can detect hybridization or PCR amplification.

ACKNOWLEDGMENTS

We acknowledge Lincoln McBride of Perkin-Elmer for his support and encouragement on this project and Mitch Winnik of the University of Toronto for helpful discussions on time-resolved fluorescence.

REFERENCES

1. Lee, L.G., C.R. Connell, and W. Bloch. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res.* **21:** 3761–3766.
2. Holland, P.M., R.D. Abramson, R. Watson, and D.H. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci.* **88:** 7276–7280.
3. Lyamichev, V., M.A.D. Brow, and J.E. Dahlberg. 1993. Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* **260:** 778–783.
4. Förster, V.Th. 1948. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys. (Leipzig)* **2:** 55–75.
5. Lakowicz, J.R. 1983. Energy transfer. In *Principles of fluorescent spectroscopy,* pp. 303–339. Plenum Press, New York, NY.
6. Stryer, L. and R.P. Haugland. 1967. Energy transfer: A spectroscopic ruler. *Proc. Natl. Acad. Sci.* **58:** 719–726.
7. Nakajima-Iijima, S., H. Hamada, P. Reddy, and T. Kakunaga. 1985. Molecular structure of the human cytoplasmic beta-actin gene: Inter-species homology of sequences in the introns. *Proc. Natl. Acad. Sci.* **82:** 6133–6137.
8. du Breuil, R.M., J.M. Patel, and B.V. Mendelow. 1993. Quantitation of β-actin-specific mRNA transcripts using xeno-competitive PCR. *PCR Methods Applic.* **3:** 57–59.
9. Livak, K.J. (unpubl.).
10. Bagwell, C.B., M.E. Munson, R.L. Christensen, and E.J. Lovett. 1994. A new homogeneous assay system for specific nucleic acid sequences: Poly-dA and poly-A detection. *Nucleic Acids Res.* **22:** 2424–2425.

Received December 20, 1994; accepted in revised form March 6, 1995.

GENOME METHODS

Real Time Quantitative PCR

Christian A. Heid,[1] Junko Stevens,[2] Kenneth J. Livak,[2] and P. Mickey Williams[1,3]

[1]BioAnalytical Technology Department, Genentech, Inc., South San Francisco, California 94080;
[2]Applied BioSystems Division of Perkin Elmer Corp., Foster City, California 94404

We have developed a novel "real time" quantitative PCR method. The method measures PCR product accumulation through a dual-labeled fluorogenic probe (i.e., TaqMan Probe). This method provides very accurate and reproducible quantitation of gene copies. Unlike other quantitative PCR methods, real-time PCR does not require post-PCR sample handling, preventing potential PCR product carry-over contamination and resulting in much faster and higher throughput assays. The real-time PCR method has a very large dynamic range of starting target molecule determination (at least five orders of magnitude). Real-time quantitative PCR is extremely accurate and less labor-intensive than current quantitative PCR methods.

Quantitative nucleic acid sequence analysis has had an important role in many fields of biological research. Measurement of gene expression (RNA) has been used extensively in monitoring biological responses to various stimuli (Tan et al. 1994; Huang et al. 1995a,b; Prud'homme et al. 1995). Quantitative gene analysis (DNA) has been used to determine the genome quantity of a particular gene, as in the case of the human *HER2* gene, which is amplified in ~30% of breast tumors (Slamon et al. 1987). Gene and genome quantitation (DNA and RNA) also have been used for analysis of human immunodeficiency virus (HIV) burden demonstrating changes in the levels of virus throughout the different phases of the disease (Connor et al. 1993; Piatak et al. 1993b; Furtado et al. 1995).

Many methods have been described for the quantitative analysis of nucleic acid sequences (both for RNA and DNA; Southern 1975; Sharp et al. 1980; Thomas 1980). Recently, PCR has proven to be a powerful tool for quantitative nucleic acid analysis. PCR and reverse transcriptase (RT)–PCR have permitted the analysis of minimal starting quantities of nucleic acid (as little as one cell equivalent). This has made possible many experiments that could not have been performed with traditional methods. Although PCR has provided a powerful tool, it is imperative that it be used properly for quantitation (Raeymaekers 1995). Many early reports of quantitative PCR and RT–PCR described quantitation of the PCR product but did not measure the initial target sequence quantity. It is essential to design proper controls for the quantitation of the initial target sequences (Ferre 1992; Clementi et al. 1993).

Researchers have developed several methods of quantitative PCR and RT–PCR. One approach measures PCR product quantity in the log phase of the reaction before the plateau (Kellogg et al. 1990; Pang et al. 1990). This method requires that each sample has equal input amounts of nucleic acid and that each sample under analysis amplifies with identical efficiency up to the point of quantitative analysis. A gene sequence (contained in all samples at relatively constant quantities, such as β-actin) can be used for sample amplification efficiency normalization. Using conventional methods of PCR detection and quantitation (gel electrophoresis or plate capture hybridization), it is extremely laborious to assure that all samples are analyzed during the log phase of the reaction (for both the target gene and the normalization gene). Another method, quantitative competitive (QC)-PCR, has been developed and is used widely for PCR quantitation. QC–PCR relies on the inclusion of an internal control competitor in each reaction (Becker-Andre 1991; Piatak et al. 1993a,b). The efficiency of each reaction is normalized to the internal competitor. A known amount of internal competitor can be

[3]Corresponding author.

added to each sample. To obtain relative quantitation, the unknown target PCR product is compared with the known competitor PCR product. Success of a quantitative competitive PCR assay relies on developing an internal control that amplifies with the same efficiency as the target molecule. The design of the competitor and the validation of amplification efficiencies require a dedicated effort. However, because QC–PCR does not require that PCR products be analyzed during the log phase of the amplification, it is the easier of the two methods to use.

Several detection systems are used for quantitative PCR and RT–PCR analysis: (1) agarose gels, (2) fluorescent labelling of PCR products and detection with laser-induced fluorescence using capillary electrophoresis (Fusco et al. 1995; Williams et al. 1996) or acrylamide gels, and (3) plate capture and sandwich probe hybridization (Mulder et al. 1994). Although these methods proved successful, each method requires post-PCR manipulations that add time to the analysis and may lead to laboratory contamination. The sample throughput of these methods is limited (with the exception of the plate capture approach), and, therefore, these methods are not well suited for uses demanding high sample throughput (i.e., screening of large numbers of biomolecules or analyzing samples for diagnostics or clinical trials).

Here we report the development of a novel assay for quantitative DNA analysis. The assay is based on the use of the 5′ nuclease assay first described by Holland et al. (1991). The method uses the 5′ nuclease activity of *Taq* polymerase to cleave a nonextendible hybridization probe during the extension phase of PCR. The approach uses dual-labeled fluorogenic hybridization probes (Lee et al. 1993; Bassler et al. 1995; Livak et al. 1995a,b). One fluorescent dye serves as a reporter [FAM (i.e., 6-carboxyfluorescein)] and its emission spectra is quenched by the second fluorescent dye, TAMRA (i.e., 6-carboxy-tetramethylrhodamine). The nuclease degradation of the hybridization probe releases the quenching of the FAM fluorescent emission, resulting in an increase in peak fluorescent emission at 518 nm. The use of a sequence detector (ABI Prism) allows measurement of fluorescent spectra of all 96 wells of the thermal cycler continuously during the PCR amplification. Therefore, the reactions are monitored in real time. The output data is described and quantitative analysis of input target DNA sequences is discussed below.

RESULTS

PCR Product Detection in Real Time

The goal was to develop a high-throughput, sensitive, and accurate gene quantitation assay for use in monitoring lipid mediated therapeutic gene delivery. A plasmid encoding human factor VIII gene sequence, pIBTM (see Methods), was used as a model therapeutic gene. The assay uses fluorescent Taqman methodology and an instrument capable of measuring fluorescence in real time (ABI Prism 7700 Sequence Detector). The Taqman reaction requires a hybridization probe labeled with two different fluorescent dyes. One dye is a reporter dye (FAM), the other is a quenching dye (TAMRA). When the probe is intact, fluorescent energy transfer occurs and the reporter dye fluorescent emission is absorbed by the quenching dye (TAMRA). During the extension phase of the PCR cycle, the fluorescent hybridization probe is cleaved by the 5′–3′ nucleolytic activity of the DNA polymerase. On cleavage of the probe, the reporter dye emission is no longer transferred efficiently to the quenching dye, resulting in an increase of the reporter dye fluorescent emission spectra. PCR primers and probes were designed for the human factor VIII sequence and human β-actin gene (as described in Methods). Optimization reactions were performed to choose the appropriate probe and magnesium concentrations yielding the highest intensity of reporter fluorescent signal without sacrificing specificity. The instrument uses a charge-coupled device (i.e., CCD camera) for measuring the fluorescent emission spectra from 500 to 650 nm. Each PCR tube was monitored sequentially for 25 msec with continuous monitoring throughout the amplification. Each tube was re-examined every 8.5 sec. Computer software was designed to examine the fluorescent intensity of both the reporter dye (FAM) and the quenching dye (TAMRA). The fluorescent intensity of the quenching dye, TAMRA, changes very little over the course of the PCR amplification (data not shown). Therefore, the intensity of TAMRA dye emission serves as an internal standard with which to normalize the reporter dye (FAM) emission variations. The software calculates a value termed ΔRn (or ΔRQ) using the following equation: $\Delta Rn = (Rn^+) - (Rn^-)$, where Rn^+ = emission intensity of reporter/emission intensity of quencher at any given time in a reaction tube, and Rn^- = emission intensity of re-

porter/emission intensity of quencher measured prior to PCR amplification in that same reaction tube. For the purpose of quantitation, the last three data points (ΔRns) collected during the extension step for each PCR cycle were analyzed. The nucleolytic degradation of the hybridization probe occurs during the extension phase of PCR, and, therefore, reporter fluorescent emission increases during this time. The three data points were averaged for each PCR cycle and the mean value for each was plotted in an "amplification plot" shown in Figure 1A. The ΔRn mean value is plotted on the *y*-axis, and time, represented by cycle number, is plotted on the *x*-axis. During the early cycles of the PCR amplification, the ΔRn value remains at base line. When sufficient hybridization probe has been cleaved by the *Taq* polymerase nuclease activity, the intensity of reporter fluorescent emission increases. Most PCR amplifications reach a plateau phase of reporter fluorescent emission if the reaction is carried out to high cycle numbers. The amplification plot is examined early in the reaction, at a point that represents the log phase of product accumulation. This is done by assigning an arbitrary threshold that is based on the variability of the base-line data. In Figure 1A, the threshold was set at 10 standard deviations above the mean of base line emission calculated from cycles 1 to 15. Once the threshold is chosen, the point at which

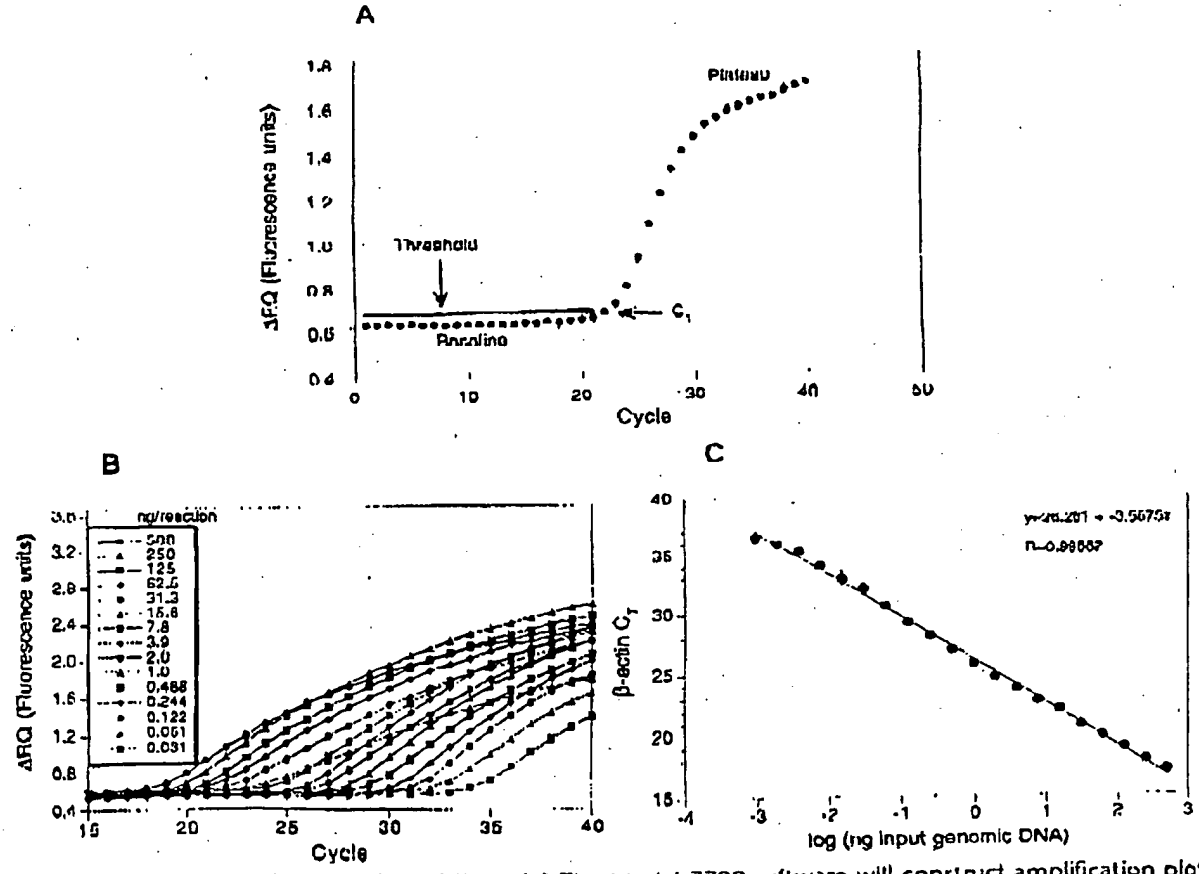


Figure 1 PCR product detection in real time. (*A*) The Model 7700 software will construct amplification plots from the extension phase fluorescent emission data collected during the PCR amplification. The standard deviation is determined from the data points collected from the base line of the amplification plot. C_T values are calculated by determining the point at which the fluorescence exceeds a threshold limit (usually 10 times the standard deviation of the base line). (*B*) Overlay of amplification plots of serially (1:2) diluted human genomic DNA samples amplified with β-actin primers. (*C*) Input DNA concentration of the samples plotted versus C_T. All

the amplification plot crosses the threshold is defined as C_T. C_T is reported as the cycle number at this point. As will be demonstrated, the C_T value is predictive of the quantity of input target.

C_T Values Provide a Quantitative Measurement of Input Target Sequences

Figure 1B shows amplification plots of 15 different PCR amplifications overlaid. The amplifications were performed on a 1:2 serial dilution of human genomic DNA. The amplified target was human β actin. The amplification plots shift to the right (to higher threshold cycles) as the input target quantity is reduced. This is expected because reactions with fewer starting copies of the target molecule require greater amplification to degrade enough probe to attain the threshold fluorescence. An arbitrary threshold of 10 standard deviations above the base line was used to determine the C_T values. Figure 1C represents the C_T values plotted versus the sample dilution value. Each dilution was amplified in triplicate PCR amplifications and plotted as mean values with error bars representing one standard deviation. The C_T values decrease linearly with increasing target quantity. Thus, C_T values can be used as a quantitative measurement of the input target number. It should be noted that the amplification plot for the 15.6-ng sample shown in Figure 1B does not reflect the same fluorescent rate of increase exhibited by most of the other samples. The 15.6-ng sample also achieves endpoint plateau at a lower fluorescent value than would be expected based on the input DNA. This phenomenon has been observed occasionally with other samples (data not shown) and may be attributable to late cycle inhibition; this hypothesis is still under investigation. It is important to note that the flattened slope and early plateau do not impact significantly the calculated C_T value as demonstrated by the fit on the line shown in Figure 1C. All triplicate amplifications resulted in very similar C_T values—the standard deviation did not exceed 0.5 for any dilution. This experiment contains a >100,000-fold range of input target molecules. Using C_T values for quantitation permits a much larger assay range than directly using total fluorescent emission intensity for quantitation. The linear range of fluorescent intensity measurement of the ABI Prism 7700 Sequence Detector ranges... ments over a very large range of relative starting target quantities.

Sample Preparation Validation

Several parameters influence the efficiency of PCR amplification: magnesium and salt concentrations, reaction conditions (i.e., time and temperature), PCR target size and composition, primer sequences, and sample purity. All of the above factors are common to a single PCR assay, except sample to sample purity. In an effort to validate the method of sample preparation for the factor VIII assay, PCR amplification reproducibility and efficiency of 10 replicate sample preparations were examined. After genomic DNA was prepared from the 10 replicate samples, the DNA was quantitated by ultraviolet spectroscopy. Amplifications were performed analyzing β-actin gene content in 100 and 25 ng of total genomic DNA. Each PCR amplification was performed in triplicate. Comparison of C_T values for each triplicate sample show minimal variation based on standard deviation and coefficient of variance (Table 1). Therefore, each of the triplicate PCR amplifications was highly reproducible, demonstrating that real time PCR using this instrumentation introduces minimal variation into the quantitative PCR analysis. Comparison of the mean C_T values of the 10 replicate sample preparations also showed minimal variability, indicating that each sample preparation yielded similar results for β-actin gene quantity. The highest C_T difference between any of the samples was 0.85 and 0.71 for the 100 and 25 ng samples, respectively. Additionally, the amplification of each sample exhibited an equivalent rate of fluorescent emission intensity change per amount of DNA target analyzed as indicated by similar slopes derived from the sample dilutions (Fig. 2). Any sample containing an excess of a PCR inhibitor would exhibit a greater measured β-actin C_T value for a given quantity of DNA. In addition, the inhibitor would be diluted along with the sample in the dilution analysis (Fig. 2), altering the expected C_T value change. Each sample amplification yielded a similar result in the analysis, demonstrating that this method of sample preparation is highly reproducible with regard to sample purity.

Quantitative Analysis of a Plasmid After

Table 1. Reproducibility of Sample Preparation Method

	100 ng				25 ng			
Sample no.	C_T	mean	standard deviation	CV	C_T	mean	standard deviation	CV
1	18.24				20.48			
	18.23				20.55			
	18.33	18.27	0.06	0.32	20.5	20.51	0.03	0.17
2	18.33				20.61			
	18.35				20.59			
	18.44	18.37	0.06	0.32	20.41	20.54	0.11	0.54
3	18.3				20.54			
	18.3				20.6			
	18.42	18.34	0.07	0.36	20.49	20.54	0.06	0.28
4	18.15				20.48			
	18.23				20.44			
	18.32	18.23	0.08	0.46	20.38	20.43	0.05	0.26
5	18.4				20.68			
	18.38				20.87			
	18.46	18.42	0.04	0.23	20.63	20.73	0.13	0.61
6	18.54				21.09			
	18.67				21.04			
	19	18.74	0.24	1.26	21.04	21.06	0.03	0.15
7	18.28				20.67			
	18.36				20.73			
	18.52	18.39	0.12	0.66	20.65	20.68	0.04	0.2
8	18.45				20.98			
	18.7				20.84			
	18.73	18.63	0.16	0.83	20.75	20.86	0.12	0.57
9	18.18				20.46			
	18.34				20.54			
	18.36	18.29	0.1	0.55	20.48	20.51	0.07	0.32
10	18.42				20.79			
	18.57				20.78			
	18.66	18.55	0.12	0.65	20.62	20.73	0.1	0.46
Mean	(1–10)	18.42	0.17	0.90		20.66	0.19	0.94

tor containing a partial cDNA for human factor VIII, pF8TM. A series of transfections was set up using a decreasing amount of the plasmid (40, 4, 0.5, and 0.1 μg). Twenty-four hours post-transfection, total DNA was purified from each flask of cells. β-Actin gene quantity was chosen as a value for normalization of genomic DNA concentration from each sample. In this experiment, β-actin gene content should remain constant relative to total genomic DNA. Figure 3 shows the result of the β-actin DNA measurement (100 ng total DNA determined by ultraviolet spectroscopy) of each sample. Each sample was analyzed in triplicate and the mean β-actin C_T values of the triplicates were plotted (error bars represent one standard deviation). The highest difference between any two sample means was 0.95 C_T. Ten nanograms of total DNA of each sample were also examined for β-actin. The results again showed that very similar amounts of genomic DNA were present; the maximum mean β actin C_T value difference was 1.0. As Figure 3 shows, the rate of β-actin C_T change between the 100 and 10-ng samples was similar (slope values range between 3.56 and −3.45). This verifies again that the method of sample preparation yields samples of identical PCR integrity (i.e., no sample contained an excessive amount of a PCR inhibitor). However, these results indicate that each sample contained slight differences in the actual amount of genomic DNA analyzed. Determination of actual genomic DNA concentration was accomplished

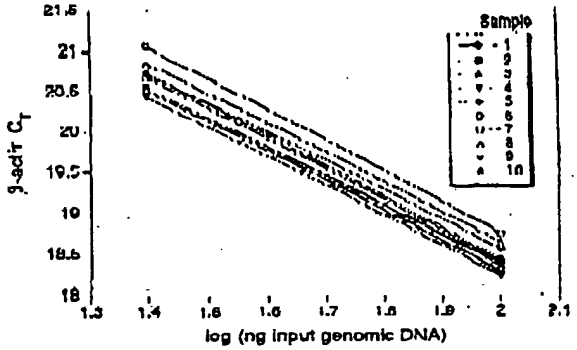


Figure 2 Sample preparation purity. The replicate samples shown in Table 1 were also amplified in tripicate using 25 ng of each DNA sample. The figure shows the input DNA concentration (100 and 25 ng) vs. C_T. In the figure, the 100 and 25 ng points for each sample are connected by a line.

by plotting the mean β-actin C_T value obtained for each 100-ng sample on a β-actin standard curve (shown in Fig. 4C). The actual genomic DNA concentration of each sample, *a*, was obtained by extrapolation to the *x*-axis.

Figure 4A shows the measured (i.e., non-normalized) quantities of factor VIII plasmid DNA (pF8TM) from each of the four transient cell transfections. Each reaction contained 100 ng of total sample DNA (as determined by UV spectroscopy). Each sample was analyzed in triplicate

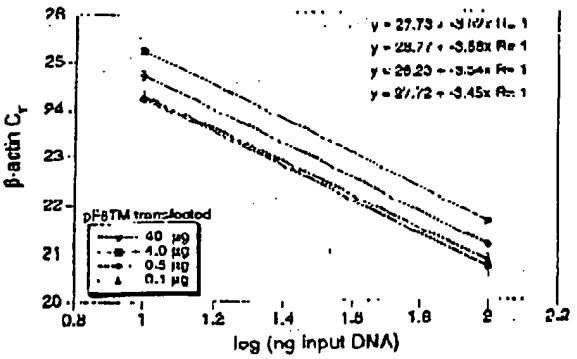


Figure 3 Analysis of transfected cell DNA quantity and purity. The DNA preparations of the four 293 cell transfections (40, 4, 0.5, and 0.1 µg of pF8TM) were analyzed for the β-actin gene. 100 and 10 ng (determined by ultraviolet spectroscopy) of each sample were amplified in triplicate. For each amount of pF8TM that was transfected, the β-actin C_T values are plotted versus the total input DNA concentration.

PCR amplifications. As shown, pF8TM purified from the 293 cells decreases (mean C_T values increase) with decreasing amounts of plasmid transfected. The mean C_T values obtained for pF8TM in Figure 4A were plotted on a standard curve comprised of serially diluted pF8TM, shown in Figure 4B. The quantity of pF8TM, *b*, found in each of the four transfections was determined by extrapolation to the *x* axis of the standard curve in Figure 4B. These uncorrected values, *b*, for pF8TM were normalized to determine the actual amount of pF8TM found per 100 ng of genomic DNA by using the equation:

$$\frac{b \times 100\text{ ng}}{a} = \text{actual pF8TM copies per 100 ng of genomic DNA}$$

where *a* = actual genomic DNA in a sample and *b* = pF8TM copies from the standard curve. The normalized quantity of pF8TM per 100 ng of genomic DNA for each of the four transfections is shown in Figure 4D. These results show that the quantity of factor VIII plasmid associated with the 293 cells, 24 hr after transfection, decreases with decreasing plasmid concentration used in the transfection. The quantity of pF8TM associated with 293 cells, after transfection with 40 µg of plasmid, was 35 pg per 100 ng genomic DNA. This results in ~520 plasmid copies per cell.

DISCUSSION

We have described a new method for quantitating gene copy numbers using real-time analysis of PCR amplifications. Real-time PCR is compatible with either of the two PCR (RT–PCR) approaches: (1) quantitative competitive where an internal competitor for each target sequence is used for normalization (data not shown) or (2) quantitative comparative PCR using a normalization gene contained within the sample (i.e., β-actin) or a "housekeeping" gene for RT–PCR. If equal amounts of nucleic acid are analyzed for each sample and if the amplification efficiency before quantitative analysis is identical for each sample, the internal control (normalization gene or competitor) should give equal signals for all samples.

The real-time PCR method offers several advantages over the other two methods currently employed (see the Introduction). First, the real-time PCR method is performed in a closed-tube system and requires no post-PCR manipulation

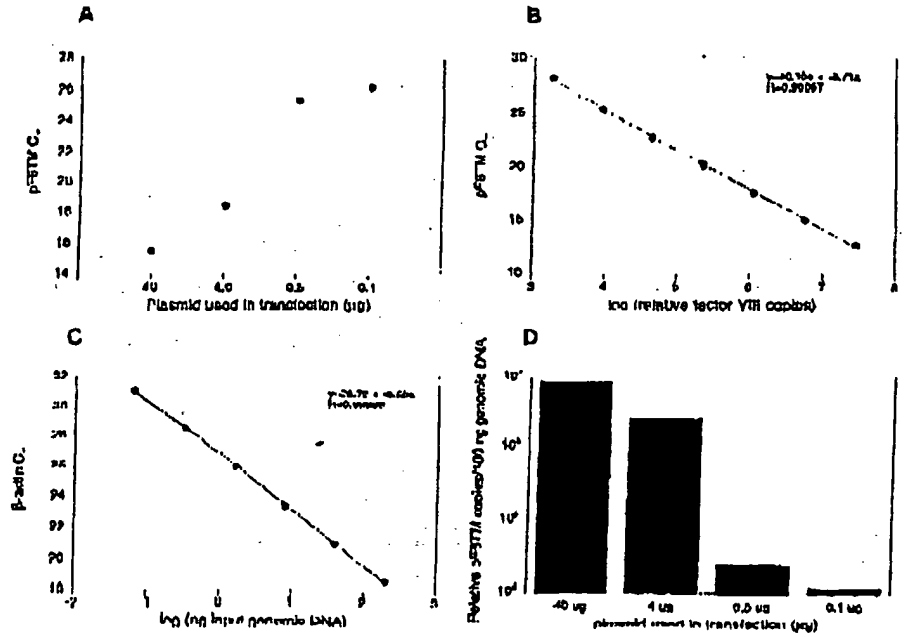


Figure 4 Quantitative analysis of pF8TM in transfected cells. (*A*) Amount of plasmid DNA used for the transfection plotted against the mean C_T value determined for pF8TM remaining 24 hr after transfection. (*B*,*C*) Standard curves of pF8TM and β-actin, respectively. pF8TM DNA (*B*) and genomic DNA (*C*) were diluted serially 1:5 before amplification with the appropriate primers. The β-actin standard curve was used to normalize the results of *A* to 100 ng of genomic DNA. (*D*) The amount of pF8TM present per 100 ng of genomic DNA.

of sample. Therefore, the potential for PCR contamination in the laboratory is reduced because amplified products can be analyzed and disposed of without opening the reaction tubes. Second, this method supports the use of a normalization gene (i.e., β-actin) for quantitative PCR or housekeeping genes for quantitative RT–PCR controls. Analysis is performed in real time during the log phase of product accumulation. Analysis during log phase permits many different genes (over a wide input target range) to be analyzed simultaneously, without concern of reaching reaction plateau at different cycles. This will make multigene analysis assays much easier to develop, because individual internal competitors will not be needed for each gene under analysis. Third, sample throughput will increase dramatically with the new method because there is no post-PCR processing time. Additionally, working in a 96-well format is highly compatible with automation technology.

The real-time PCR method is highly reproducible. Replicate amplifications can be analyzed for each sample minimizing potential error. The system allows for a very large assay dynamic range (approaching 1,000,000-fold starting target). Using a standard curve for the target of interest, relative copy number values can be determined for any unknown sample. Fluorescent threshold values, C_T, correlate linearly with relative DNA copy numbers. Real time quantitative RT-PCR methodology (Gibson et al., this issue) has also been developed. Finally, real time quantitative PCR methodology can be used to develop high-throughput screening assays for a variety of applications [quantitative gene expression (RT–PCR), gene copy assays (Her2, HIV, etc.), genotyping (knockout mouse analysis), and Immuno-PCR].

Real-time PCR may also be performed using intercalating dyes (Higuchi et al. 1992) such as ethidium bromide. The fluorogenic probe method offers a major advantage over intercalating dyes—greater specificity (i.e., primer dimers and nonspecific PCR products are not detected).

METHODS

Generation of a Plasmid Containing a Partial cDNA for Human Factor VIII

Total RNA was harvested (RNAzol B from Tel Test, Inc., Friendswood, TX) from cells transfected with a factor VIII expression vector, pCIS2.8c251) (Eaton et al. 1986; Gorman et al. 1990). A factor VIII partial cDNA sequence was generated by RT PCR [GeneAmp RNA PCR Kit (part N808-0179, PE Applied Biosystems, Foster City, CA)] using the PCR primers F8for and F8rev (primer sequences are shown below). The amplicon was reamplified using modified F8for and F8rev primers (appended with *Bam*HI and *Hin*dIII restriction site sequences at the 5' end) and cloned into pGEM-3Z (Promega Corp., Madison, WI). The resulting clone, pF8TM, was used for transient transfection of 293 cells.

Amplification of Target DNA and Detection of Amplicon Factor VIII Plasmid DNA

(pF8TM) was amplified with the primers F8for 5'-CCCGTTCCCAAGAGTTGACTTTCTC-3' and F8rev 5'-AAACCTTCACCCCTGGATGCTTACG-3'. The reaction produced a 422-bp PCR product. The forward primer was designed to recognize a unique sequence found in the 5' untranslated region of the parent pCIS2.8c251) plasmid and therefore does not recognize and amplify the human factor VIII gene. Primers were chosen with the assistance of the computer program Oligo 4.0 (National Biosciences, Inc., Plymouth, MN). The human β-actin gene was amplified with the primers β-actin forward primer 5'-TCACCCACACTGTGCCCATCTACGA-3' and β-actin reverse primer 5'-CAGCGGAACCGCTCATTGCCAATGG-3'. The reaction produced a 295-bp PCR product.

Amplification reactions (50 μl) contained a DNA sample, 10× PCR Buffer II (5 μl), 200 μM dATP, dCTP, dGTP, and 400 μM dUTP, 4 mM $MgCl_2$, 1.25 Units AmpliTaq DNA polymerase, 0.5 unit AmpErase uracil *N*-glycosylase (UNG), 60 pmole of each factor VIII primer, and 15 pmole of each β actin primer. The reactions also contained one of the following detection probes (100 nM each): F8probe 5'(FAM)AGCTCTCCCACCTGCTTTCTTTCTGTGCCTT(TAMRA)p 3' and β-actin probe 5' (FAM)ATGCCCX(TAMRA)CCCCCATGCCATCp-3' where p indicates phosphorylation and X indicates a linker arm nucleotide. Reaction tubes were MicroAmp Optical Tubes (part number N801 0933, Perkin Elmer) that were frosted (at Perkin Elmer) to prevent light from reflecting. Tube caps were similar to MicroAmp Caps but specially designed to prevent light scattering. All of the PCR consumables were supplied by PE Applied Biosystems (Foster City, CA) except the factor VIII primers, which were synthesized at Genentech, Inc. (South San Francisco, CA). Probes were designed using the Oligo 4.0 software, following guidelines suggested in the Model 7700 Sequence Detector instrument manual. Briefly, probe T_m should be at least 5°C higher than the annealing temperature used during thermal cycling; primers should not form stable duplexes with the probe.

The thermal cycling conditions included 2 min at 50°C and 10 min at 95°C. Thermal cycling proceeded with 40 cycles of 95°C for 15 sec and 60°C for 1 min. All reactions were performed in the Model 7700 Sequence Detector (PE Applied Biosystems), which contains a GeneAmp PCR System 9600. Reaction conditions were programmed on a Power Macintosh 7100 (Apple Computer, Santa Clara, CA) linked directly to the Model 7700 Sequence Detector. Analysis of data was also performed on the Macintosh computer. Collection and analysis software was developed at PE Applied Biosystems.

Transfection of Cells with Factor VIII Construct

Four T175 flasks of 293 cells (ATCC CRL 1573), a human fetal kidney suspension cell line, were grown to 80% confluency and transfected pF8TM. Cells were grown in the following media: 50% HAMS F12 without GHT, 50% low glucose Dulbecco's modified Eagle medium (DMEM) without glycine with sodium bicarbonate, 10% fetal bovine serum, 2 mM L-glutamine, and 1% penicillin-streptomycin. The media was changed 30 min before the transfection. pF8TM DNA amounts of 40, 4, 0.5, and 0.1 μg were added to 1.5 ml of a solution containing 0.125 M $CaCl_2$ and 1× HEPES. The four mixtures were left at room temperature for 10 min and then added dropwise to the cells. The flasks were incubated at 37°C and 5% CO_2 for 24 hr, washed with PBS, and resuspended in PBS. The resuspended cells were divided into aliquots and DNA was extracted immediately using the QIAamp Blood Kit (Qiagen, Chatsworth, CA). DNA was eluted into 200 μl of 20 mM Tris-HCl at pH 8.0.

ACKNOWLEDGMENTS

We thank Genentech's DNA Synthesis Group for primer synthesis and Genentech's Graphics Group for assistance with the figures.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bassler, H.A., S.J. Flood, K.J. Livak, J. Marmaro, R. Knorr, and C.A. Batt. 1995. Use of a fluorogenic probe in a PCR-based assay for the detection of Listeria monocytogenes. *App. Environ. Microbiol.* 61: 3724–3728.

Becker-Andre, M. 1991. Quantitative evaluation of mRNA levels. *Meth. Mol. Cell. Biol.* 2: 189–201.

Clementi, M., S. Menzo, P. Bagnarelli, A. Manzin, A. Valenza, and P.E. Varaldo. 1993. Quantitative PCR and RT-PCR in virology. [Review]. *PCR Methods Applic.* 2: 191–196.

Connor, R.I., H. Mohri, Y. Cao, and D.D. Ho. 1993. Increased viral burden and cytopathicity correlate temporally with CD4+ T-lymphocyte decline and clinical progression in human immunodeficiency virus type 1-infected individuals. *J. Virol.* 67: 1772–1777.

Eaton, D.L., W.I. Wood, D. Eaton, P.E. Hass, P.

Vehar, and G. Gorman. 1986. Construction and characterization of an active factor VIII variant lacking the central one third of the molecule. *Biochemistry* 25: 8343–8347.

Fasco, M.J., C.P. Treanor, S. Spivack, H.L. Rigge, and L.S. Kaminsky. 1995. Quantitative RNA-polymerase chain reaction-DNA analysis by capillary electrophoresis and laser-induced fluorescence. *Anal. Biochem.* 224: 140–147.

Ferre, F. 1992. Quantitative or semi-quantitative PCR: Reality versus myth. *PCR Methods Applic.* 2: 1–9.

Furtado, M.R., L.A. Kingsley, and S.M. Wolinsky. 1995. Changes in the viral mRNA expression pattern correlate with a rapid rate of CD4+ T-cell number decline in human immunodeficiency virus type 1-infected individuals. *J. Virol.* 69: 2092–2100.

Gibson, U.E.M., C.A. Heid, and P.M. Williams. 1996. A novel method for real time quantitative competitive RT–PCR. *Genome Res.* (this issue).

Gorman, C.M., D.R. Gies, and G. McCray. 1990. Transient production of proteins using an adenovirus transformed cell line. *DNA Prot. Engin. Tech.* 2: 3–10.

Higuchi, R., G. Dollinger, P.S. Walsh, and R. Griffith. 1992. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology* 10: 413–417.

Holland, P.M., R.D. Abramson, R. Watson, and D.H. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5′—3′ exonuclease activity of Thermus aquaticus DNA polymerase. *Proc. Natl. Acad. Sci.* 88: 7276–7280.

Huang, S.K., H.Q. Xiao, T.J. Kleine, G. Pacciotti, D.G. Marsh, L.M. Lichtenstein, and M.C. Liu. 1995a. IL-13 expression at the sites of allergen challenge in patients with asthma. *J. Immun.* 155: 2688–2694.

Huang, S.K., M. Yi, E. Palmer, and D.G. Marsh. 1995b. A dominant T cell receptor beta-chain in response to a short ragweed allergen, Amb a 5. *J. Immun.* 154: 6157–6162.

Kellogg, D.E., J.J. Sninsky, and S. Kowk. 1990. Quantitation of HIV-1 proviral DNA relative to cellular DNA by the polymerase chain reaction. *Anal. Biochem.* 189: 202–208.

Lee, L.G., C.R. Connell, and W. Bloch. 1993. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res.* 21: 3761–3766.

Livak, K.J., S.J. Flood, J. Marmaro, W. Giusti, and K. Deetz. 1995a. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Applic.* 4: 357–362.

Livak, K.J., J. Marmaro, and J.A. Todd. 1995b. Towards fully automated genome-wide polymorphism screening [Letter]. *Nature Genet.* 9: 341–342.

Mulder, J., N. McKinney, C. Christopherson, J. Sninsky, L. Greenfield, and S. Kwok. 1994. Rapid and simple PCR assay for quantitation of human immunodeficiency virus type 1 RNA in plasma: Application to acute retroviral infection. *J. Clin. Microbiol.* 32: 292–300.

Pang, S., Y. Koyanagi, S. Miles, C. Wiley, H.V. Vinters, and I.S. Chen. 1990. High levels of unintegrated HIV-1 DNA in brain tissue of AIDS dementia patients. *Nature* 343: 85–89.

Piatak, M.J., K.C. Luk, B. Williams, and J.D. Lifson. 1993a. Quantitative competitive polymerase chain reaction for accurate quantitation of HIV DNA and RNA species. *BioTechniques* 14: 70–81.

Piatak, M.J., M.S. Saag, L.C. Yang, S.J. Clark, J.C. Kappes, K.C. Luk, B.H. Hahn, G.M. Shaw, and J.D. Lifson. 1993b. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR [see Comments]. *Science* 259: 1749–1754.

Prud'homme, G.J., D.H. Kono, and A.N. Theofilopoulos. 1995. Quantitative polymerase chain reaction analysis reveals marked overexpression of interleukin-1 beta, interleukin-1 and interferon-gamma mRNA in the lymph nodes of lupus-prone mice. *Mol. Immunol.* 32: 495–503.

Raeymaekers, L. 1995. A commentary on the practical applications of competitive PCR. *Genome Res.* 5: 91–94.

Sharp, P.A., A.J. Berk, and S.M. Berget. 1980. Transcription maps of adenovirus. *Methods Enzymol.* 65: 750–768.

Slamon, D.J., G.M. Clark, S.G. Wong, W.J. Levin, A. Ullrich, and W.L. McGuire. 1987. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235: 177–182.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98: 503–517.

Tan, X., X. Sun, G.P. Gonzalez, and W. Hsueh. 1994. PAF and TNF increase the precursor of NF-kappa B p50 mRNA in mouse intestine: Quantitative analysis by competitive PCR. *Biochim. Biophys. Acta* 1215: 157–162.

Thomas, P.S. 1980. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proc. Natl. Acad. Sci.* 77: 5201–5205.

Williams, S., C. Schwer, A. Krishnarao, C. Heid, B. Karger, and P.M. Williams. 1996. Quantitative competitive PCR: Analysis of amplified products of the HIV-1 gag gene by capillary electrophoresis with laser induced fluorescence detection. *Anal. Biochem.* (in press).

Received June 3, 1996; accepted in revised form July 29, 1996.

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14717–14722, December 1998
Cell Biology, Medical Sciences.

WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors

DIANE PENNICA*†, TODD A. SWANSON*, JAMES W. WELSH*, MARGARET A. ROY‡, DAVID A. LAWRENCE*, JAMES LEE‡, JENNIFER BRUSH‡, LISA A. TANEYHILL§, BETHANNE DEUEL‡, MICHAEL LEW¶, COLIN WATANABE‖, ROBERT L. COHEN*, MONA F. MELHEM**, GENE G. FINLEY**, PHIL QUIRKE††, AUDREY D. GODDARD‡, KENNETH J. HILLAN¶, AUSTIN L. GURNEY‡, DAVID BOTSTEIN‡,‡‡, AND ARNOLD J. LEVINE§

Departments of *Molecular Oncology, ‡Molecular Biology, ¶Scientific Computing, and ¶Pathology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080; **University of Pittsburgh School of Medicine, Veterans Administration Medical Center, Pittsburgh, PA 15240; ††University of Leeds, Leeds, LS29JT United Kingdom; ‡‡Department of Genetics, Stanford University, Palo Alto, CA 94305; and §Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Contributed by David Botstein and Arnold J. Levine, October 21, 1998

ABSTRACT **Wnt family members are critical to many developmental processes, and components of the Wnt signaling pathway have been linked to tumorigenesis in familial and sporadic colon carcinomas. Here we report the identification of two genes, *WISP-1* and *WISP-2*, that are up-regulated in the mouse mammary epithelial cell line C57MG transformed by Wnt-1, but not by Wnt-4. Together with a third related gene, *WISP-3*, these proteins define a subfamily of the connective tissue growth factor family. Two distinct systems demonstrated *WISP* induction to be associated with the expression of Wnt-1. These included (*i*) C57MG cells infected with a Wnt-1 retroviral vector or expressing Wnt-1 under the control of a tetracyline repressible promoter, and (*ii*) Wnt-1 transgenic mice. The *WISP-1* gene was localized to human chromosome 8q24.1–8q24.3. *WISP-1* genomic DNA was amplified in colon cancer cell lines and in human colon tumors and its RNA overexpressed (2- to >30-fold) in 84% of the tumors examined compared with patient-matched normal mucosa. *WISP-3* mapped to chromosome 6q22–6q23 and also was overexpressed (4- to >40-fold) in 63% of the colon tumors analyzed. In contrast, *WISP-2* mapped to human chromosome 20q12–20q13 and its DNA was amplified, but RNA expression was reduced (2- to >30-fold) in 79% of the tumors. These results suggest that the *WISP* genes may be downstream of Wnt-1 signaling and that aberrant levels of *WISP* expression in colon cancer may play a role in colon tumorigenesis.**

Wnt-1 is a member of an expanding family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes such as the control of cell proliferation, adhesion, cell polarity, and the establishment of cell fates (1, 2). Wnt-1 originally was identified as an oncogene activated by the insertion of mouse mammary tumor virus in virus-induced mammary adenocarcinomas (3, 4). Although Wnt-1 is not expressed in the normal mammary gland, expression of Wnt-1 in transgenic mice causes mammary tumors (5).

In mammalian cells, Wnt family members initiate signaling by binding to the seven-transmembrane spanning Frizzled receptors and recruiting the cytoplasmic protein Dishevelled (Dsh) to the cell membrane (1, 2, 6). Dsh then inhibits the kinase activity of the normally constitutively active glycogen synthase kinase-3β (GSK-3β) resulting in an increase in β-catenin levels. Stabilized β-catenin interacts with the transcription factor TCF/Lef1, forming a complex that appears in the nucleus and binds TCF/Lef1 target DNA elements to activate transcription (7, 8). Other experiments suggest that the adenomatous polyposis coli (APC) tumor suppressor gene also plays an important role in Wnt signaling by regulating β-catenin levels (9). APC is phosphorylated by GSK-3β, binds to β-catenin, and facilitates its degradation. Mutations in either APC or β-catenin have been associated with colon carcinomas and melanomas, suggesting these mutations contribute to the development of these types of cancer, implicating the Wnt pathway in tumorigenesis (1).

Although much has been learned about the Wnt signaling pathway over the past several years, only a few of the transcriptionally activated downstream components activated by Wnt have been characterized. Those that have been described cannot account for all of the diverse functions attributed to Wnt signaling. Among the candidate Wnt target genes are those encoding the nodal-related 3 gene, *Xnr3*, a member of the transforming growth factor (TGF)-β superfamily, and the homeobox genes, *engrailed*, *goosecoid*, *twin* (*Xtwn*), and *siamois* (2). A recent report also identifies c-*myc* as a target gene of the Wnt signaling pathway (10).

To identify additional downstream genes in the Wnt signaling pathway that are relevant to the transformed cell phenotype, we used a PCR-based cDNA subtraction strategy, suppression subtractive hybridization (SSH) (11), using RNA isolated from C57MG mouse mammary epithelial cells and C57MG cells stably transformed by a Wnt-1 retrovirus. Overexpression of Wnt-1 in this cell line is sufficient to induce a partially transformed phenotype, characterized by elongated and refractile cells that lose contact inhibition and form a multilayered array (12, 13). We reasoned that genes differentially expressed between these two cell lines might contribute to the transformed phenotype.

In this paper, we describe the cloning and characterization of two genes up-regulated in Wnt-1 transformed cells, *WISP-1* and *WISP-2*, and a third related gene, *WISP-3*. The *WISP* genes are members of the CCN family of growth factors, which includes connective tissue growth factor (CTGF), Cyr61, and *nov*, a family not previously linked to Wnt signaling.

MATERIALS AND METHODS

SSH. SSH was performed by using the PCR-Select cDNA Subtraction Kit (CLONTECH). Tester double-stranded

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: TGF, transforming growth factor; CTGF, connective tissue growth factor; SSH, suppression subtractive hybridization; VWC, von Willebrand factor type C module.

Data deposition: The sequences reported in this paper have been deposited in the Genbank database (accession nos. AF100777, AF100778, AF100779, AF100780, and AF100781).

†To whom reprint requests should be addressed. e-mail: diane@gene.com.

cDNA was synthesized from 2 μg of poly(A)$^+$ RNA isolated from the C57MG/Wnt-1 cell line and driver cDNA from 2 μg of poly(A)$^+$ RNA from the parent C57MG cells. The subtracted cDNA library was subcloned into a pGEM-T vector for further analysis.

cDNA Library Screening. Clones encoding full-length mouse *WISP-1* were isolated by screening a λgt10 mouse embryo cDNA library (CLONTECH) with a 70-bp probe from the original partial clone 568 sequence corresponding to amino acids 128–169. Clones encoding full-length human *WISP-1* were isolated by screening λgt10 lung and fetal kidney cDNA libraries with the same probe at low stringency. Clones encoding full-length mouse and human *WISP-2* were isolated by screening a C57MG/Wnt-1 or human fetal lung cDNA library with a probe corresponding to nucleotides 1463–1512. Full-length cDNAs encoding *WISP-3* were cloned from human bone marrow and fetal kidney libraries.

Expression of Human *WISP* RNA. PCR amplification of first-strand cDNA was performed with human Multiple Tissue cDNA panels (CLONTECH) and 300 μM of each dNTP at 94°C for 1 sec, 62°C for 30 sec, 72°C for 1 min, for 22–32 cycles. *WISP* and glyceraldehyde-3-phosphate dehydrogenase primer sequences are available on request.

***In Situ* Hybridization.** ^{33}P-labeled sense and antisense riboprobes were transcribed from an 897-bp PCR product corresponding to nucleotides 601–1440 of mouse *WISP-1* or a 294-bp PCR product corresponding to nucleotides 82–375 of mouse *WISP-2*. All tissues were processed as described (40).

Radiation Hybrid Mapping. Genomic DNA from each hybrid in the Stanford G3 and Genebridge4 Radiation Hybrid Panels (Research Genetics, Huntsville, AL) and human and hamster control DNAs were PCR-amplified, and the results were submitted to the Stanford or Massachusetts Institute of Technology web servers.

Cell Lines, Tumors, and Mucosa Specimens. Tissue specimens were obtained from the Department of Pathology (University of Pittsburgh) for patients undergoing colon resection and from the University of Leeds, United Kingdom. Genomic DNA was isolated (Qiagen) from the pooled blood of 10 normal human donors, surgical specimens, and the following ATCC human cell lines: SW480, COLO 320DM, HT-29, WiDr, and SW403 (colon adenocarcinomas), SW620 (lymph node metastasis, colon adenocarcinoma), HCT 116 (colon carcinoma), SK-CO-1 (colon adenocarcinoma, ascites), and HM7 (a variant of ATCC colon adenocarcinoma cell line LS 174T). DNA concentration was determined by using Hoechst dye 33258 intercalation fluorimetry. Total RNA was prepared by homogenization in 7 M GuSCN followed by centrifugation over CsCl cushions or prepared by using RNAzol.

Gene Amplification and RNA Expression Analysis. Relative gene amplification and RNA expression of *WISP*s and c-*myc* in the cell lines, colorectal tumors, and normal mucosa were determined by quantitative PCR. Gene-specific primers and fluorogenic probes (sequences available on request) were designed and used to amplify and quantitate the genes. The relative gene copy number was derived by using the formula $2^{(\Delta ct)}$ where ΔCt represents the difference in amplification cycles required to detect the *WISP* genes in peripheral blood lymphocyte DNA compared with colon tumor DNA or colon tumor RNA compared with normal mucosal RNA. The ∂-method was used for calculation of the SE of the gene copy number or RNA expression level. The *WISP*-specific signal was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase housekeeping gene. All TaqMan assay reagents were obtained from Perkin–Elmer Applied Biosystems.

RESULTS

Isolation of *WISP-1* and *WISP-2* by SSH. To identify Wnt-1-inducible genes, we used the technique of SSH using the mouse mammary epithelial cell line C57MG and C57MG cells that stably express Wnt-1 (11). Candidate differentially expressed cDNAs (1,384 total) were sequenced. Thirty-nine percent of the sequences matched known genes or homologues, 32% matched expressed sequence tags, and 29% had no match. To confirm that the transcript was differentially expressed, semiquantitative reverse transcription–PCR and Northern analysis were performed by using mRNA from the C57MG and C57MG/Wnt-1 cells.

Two of the cDNAs, *WISP-1* and *WISP-2*, were differentially expressed, being induced in the C57MG/Wnt-1 cell line, but not in the parent C57MG cells or C57MG cells overexpressing Wnt-4 (Fig. 1 *A* and *B*). Wnt-4, unlike Wnt-1, does not induce the morphological transformation of C57MG cells and has no effect on β-catenin levels (13, 14). Expression of *WISP-1* was up-regulated approximately 3-fold in the C57MG/Wnt-1 cell line and *WISP-2* by approximately 5-fold by both Northern analysis and reverse transcription–PCR.

An independent, but similar, system was used to examine *WISP* expression after Wnt-1 induction. C57MG cells expressing the *Wnt-1* gene under the control of a tetracycline-repressible promoter produce low amounts of Wnt-1 in the repressed state but show a strong induction of *Wnt-1* mRNA and protein within 24 hr after tetracycline removal (8). The levels of Wnt-1 and *WISP* RNA isolated from these cells at various times after tetracycline removal were assessed by quantitative PCR. Strong induction of Wnt-1 mRNA was seen as early as 10 hr after tetracycline removal. Induction of *WISP* mRNA (2- to 6-fold) was seen at 48 and 72 hr (data not shown). These data support our previous observations that show that *WISP* induction is correlated with Wnt-1 expression. Because the induction is slow, occurring after approximately 48 hr, the induction of *WISP*s may be an indirect response to Wnt-1 signaling.

cDNA clones of human *WISP-1* were isolated and the sequence compared with mouse *WISP-1*. The cDNA sequences of mouse and human *WISP-1* were 1,766 and 2,830 bp in length, respectively, and encode proteins of 367 aa, with predicted relative molecular masses of ≈40,000 (M_r 40 K). Both have hydrophobic N-terminal signal sequences, 38 conserved cysteine residues, and four potential N-linked glycosylation sites and are 84% identical (Fig. 2*A*).

Full-length cDNA clones of mouse and human *WISP-2* were 1,734 and 1,293 bp in length, respectively, and encode proteins of 251 and 250 aa, respectively, with predicted relative molecular masses of ≈27,000 (M_r 27 K) (Fig. 2*B*). Mouse and human *WISP-2* are 73% identical. Human *WISP-2* has no potential N-linked glycosylation sites, and mouse *WISP-2* has one at

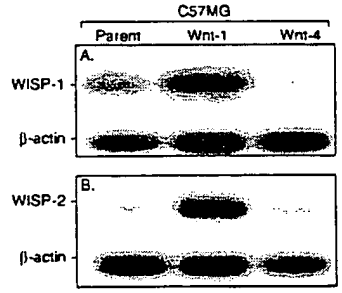


FIG. 1. *WISP-1* and *WISP-2* are induced by Wnt-1, but not Wnt-4, expression in C57MG cells. Northern analysis of *WISP-1* (*A*) and *WISP-2* (*B*) expression in C57MG, C57MG/Wnt-1, and C57MG/Wnt-4 cells. Poly(A)$^+$ RNA (2 μg) was subjected to Northern blot analysis and hybridized with a 70-bp mouse *WISP-1*-specific probe (amino acids 278–300) or a 190-bp *WISP-2*-specific probe (nucleotides 1438–1627) in the 3′ untranslated region. Blots were rehybridized with human β-actin probe.

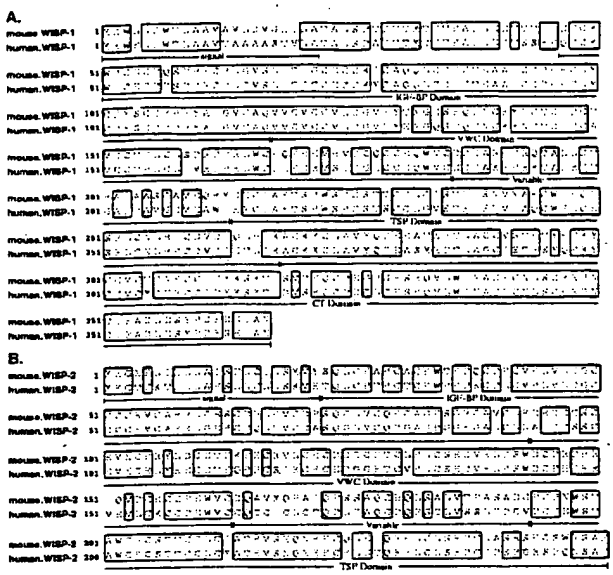


FIG. 2. Encoded amino acid sequence alignment of mouse and human *WISP-1* (*A*) and mouse and human *WISP-2* (*B*). The potential signal sequence, insulin-like growth factor-binding protein (IGF-BP), VWC, thrombospondin (TSP), and C-terminal (CT) domains are underlined.

position 197. *WISP-2* has 28 cysteine residues that are conserved among the 38 cysteines found in *WISP-1*.

Identification of *WISP-3*. To search for related proteins, we screened expressed sequence tag (EST) databases with the WISP-1 protein sequence and identified several ESTs as potentially related sequences. We identified a homologous protein that we have called WISP-3. A full-length human *WISP-3* cDNA of 1,371 bp was isolated corresponding to those ESTs that encode a 354-aa protein with a predicted molecular mass of 39,293. WISP-3 has two potential N-linked glycosylation sites and 36 cysteine residues. An alignment of the three human WISP proteins shows that WISP-1 and WISP-3 are the most similar (42% identity), whereas WISP-2 has 37% identity with WISP-1 and 32% identity with WISP-3 (Fig. 3*A*).

***WISPs* Are Homologous to the CTGF Family of Proteins.** Human *WISP-1*, *WISP-2*, and *WISP-3* are novel sequences; however, mouse *WISP-1* is the same as the recently identified *Elm1* gene. *Elm1* is expressed in low, but not high, metastatic mouse melanoma cells, and suppresses the *in vivo* growth and metastatic potential of K-1735 mouse melanoma cells (15). Human and mouse *WISP-2* are homologous to the recently described rat gene, *rCop-1* (16). Significant homology (36–44%) was seen to the CCN family of growth factors. This family includes three members, CTGF, Cyr61, and the protooncogene *nov*. CTGF is a chemotactic and mitogenic factor for fibroblasts that is implicated in wound healing and fibrotic disorders and is induced by TGF-β (17). Cyr61 is an extracellular matrix signaling molecule that promotes cell adhesion, proliferation, migration, angiogenesis, and tumor growth (18, 19). *nov* (nephroblastoma overexpressed) is an immediate early gene associated with quiescence and found altered in Wilms tumors (20). The proteins of the CCN family share functional, but not sequence, similarity to Wnt-1. All are secreted, cysteine-rich heparin binding glycoproteins that associate with the cell surface and extracellular matrix.

WISP proteins exhibit the modular architecture of the CCN family, characterized by four conserved cysteine-rich domains (Fig. 3*B*) (21). The N-terminal domain, which includes the first 12 cysteine residues, contains a consensus sequence (GCGCCXXC) conserved in most insulin-like growth factor (IGF)-binding proteins (BP). This sequence is conserved in WISP-2 and WISP-3, whereas WISP-1 has a glutamine in the third position instead of a glycine. CTGF recently has been shown to specifically bind IGF (22) and a truncated *nov* protein lacking the IGF-BP domain is oncogenic (23). The von Willebrand factor type C module (VWC), also found in certain collagens and mucins, covers the next 10 cysteine residues, and is thought to participate in protein complex formation and oligomerization (24). The VWC domain of WISP-3 differs from all CCN family members described previously, in that it contains only six of the 10 cysteine residues (Fig. 3 *A* and *B*). A short variable region follows the VWC domain. The third module, the thrombospondin (TSP) domain is involved in binding to sulfated glycoconjugates and contains six cysteine residues and a conserved WSxCSxxCG motif first identified in thrombospondin (25). The C-terminal (CT) module containing the remaining 10 cysteines is thought to be involved in dimerization and receptor binding (26). The CT domain is present in all CCN family members described to date but is absent in WISP-2 (Fig. 3 *A* and *B*). The existence of a putative signal sequence and the absence of a transmembrane domain suggest that WISPs are secreted proteins, an observation supported by an analysis of their expression and secretion from mammalian cell and baculovirus cultures (data not shown).

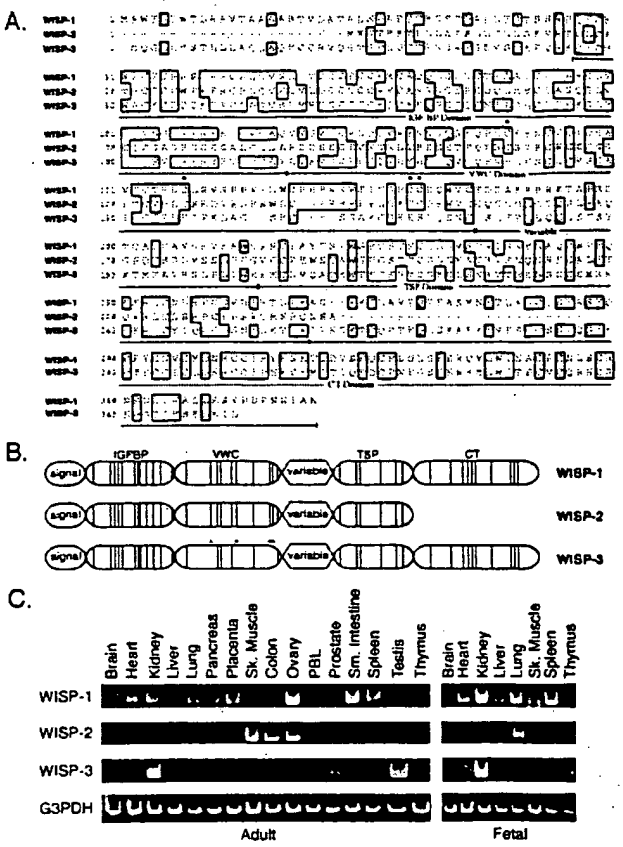


FIG. 3. (*A*) Encoded amino acid sequence alignment of human WISPs. The cysteine residues of WISP-1 and WISP-2 that are not present in WISP-3 are indicated with a dot. (*B*) Schematic representation of the WISP proteins showing the domain structure and cysteine residues (vertical lines). The four cysteine residues in the VWC domain that are absent in WISP-3 are indicated with a dot. (*C*) Expression of *WISP* mRNA in human tissues. PCR was performed on human multiple-tissue cDNA panels (CLONTECH) from the indicated adult and fetal tissues.

Expression of *WISP* mRNA in Human Tissues. Tissue-specific expression of human *WISP*s was characterized by PCR

analysis on adult and fetal multiple tissue cDNA panels. *WISP-1* expression was seen in the adult heart, kidney, lung, pancreas, placenta, ovary, small intestine, and spleen (Fig. 3*C*). Little or no expression was detected in the brain, liver, skeletal muscle, colon, peripheral blood leukocytes, prostate, testis, or thymus. *WISP-2* had a more restricted tissue expression and was detected in adult skeletal muscle, colon, ovary, and fetal lung. Predominant expression of *WISP-3* was seen in adult kidney and testis and fetal kidney. Lower levels of *WISP-3* expression were detected in placenta, ovary, prostate, and small intestine.

***In Situ* Localization of *WISP-1* and *WISP-2*.** Expression of *WISP-1* and *WISP-2* was assessed by *in situ* hybridization in mammary tumors from Wnt-1 transgenic mice. Strong expression of *WISP-1* was observed in stromal fibroblasts lying within the fibrovascular tumor stroma (Fig. 4 *A–D*). However, low-level *WISP-1* expression also was observed focally within tumor cells (data not shown). No expression was observed in normal breast. Like *WISP-1*, *WISP-2* expression also was seen in the tumor stroma in breast tumors from Wnt-1 transgenic animals (Fig. 4 *E–H*). However, *WISP-2* expression in the stroma was in spindle-shaped cells adjacent to capillary vessels, whereas the predominant cell type expressing *WISP-1* was the stromal fibroblasts.

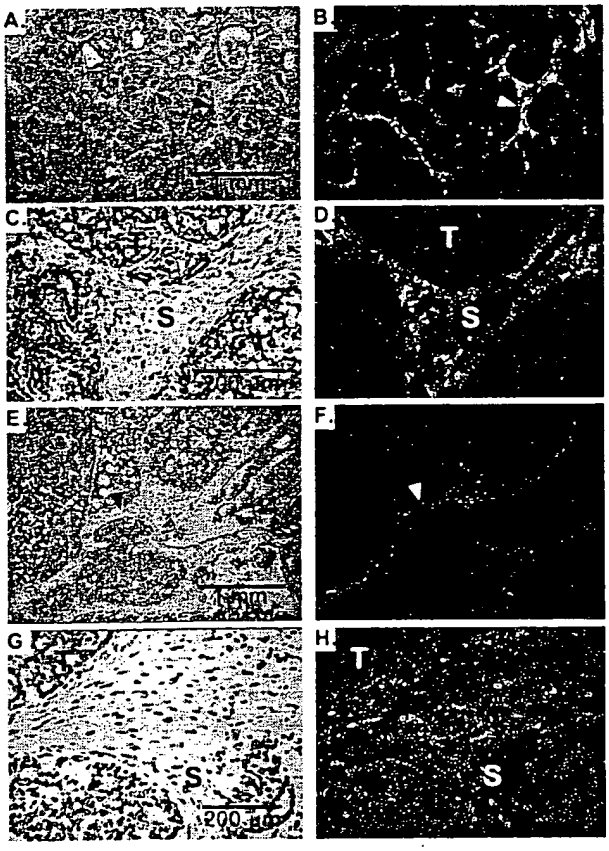


FIG. 4. (*A*, *C*, *E*, and *G*) Representative hematoxylin/eosin-stained images from breast tumors in Wnt-1 transgenic mice. The corresponding dark-field images showing *WISP-1* expression are shown in *B* and *D*. The tumor is a moderately well-differentiated adenocarcinoma showing evidence of adenoid cystic change. At low power (*A* and *B*), expression of *WISP-1* is seen in the delicate branching fibrovascular tumor stroma (arrowhead). At higher magnification, expression is seen in the stromal(s) fibroblasts (*C* and *D*), and tumor cells are negative. Focal expression of *WISP-1*, however, was observed in tumor cells in some areas. Images of *WISP-2* expression are shown in *E–H*. At low power (*E* and *F*), expression of WISP-2 is seen in cells lying within the fibrovascular tumor stroma. At higher magnification, these cells appeared to be adjacent to capillary vessels whereas tumor cells are negative (*G* and *H*).

Chromosome Localization of the *WISP* Genes. The chromosomal location of the human *WISP* genes was determined by radiation hybrid mapping panels. *WISP-1* is approximately 3.48 cR from the meiotic marker AFM259xc5 [logarithm of odds (lod) score 16.31] on chromosome 8q24.1 to 8q24.3, in the same region as the human locus of the *nov*H family member (27) and roughly 4 Mbs distal to c-*myc* (28). Preliminary fine mapping indicates that *WISP-1* is located near D8S1712 STS. *WISP-2* is linked to the marker SHGC-33922 (lod = 1,000) on chromosome 20q12–20q13.1. Human *WISP-3* mapped to chromosome 6q22–6q23 and is linked to the marker AFM211ze5 (lod = 1,000). *WISP-3* is approximately 18 Mbs proximal to CTGF and 23 Mbs proximal to the human cellular oncogene *MYB* (27, 29).

Amplification and Aberrant Expression of *WISPs* in Human Colon Tumors. Amplification of protooncogenes is seen in many human tumors and has etiological and prognostic significance. For example, in a variety of tumor types, c-*myc* amplification has been associated with malignant progression and poor prognosis (30). Because *WISP-1* resides in the same general chromosomal location (8q24) as c-*myc*, we asked whether it was a target of gene amplification, and, if so, whether this amplification was independent of the c-*myc* locus. Genomic DNA from human colon cancer cell lines was assessed by quantitative PCR and Southern blot analysis. (Fig. 5 *A* and *B*). Both methods detected similar degrees of *WISP-1* amplification. Most cell lines showed significant (2- to 4-fold) amplification, with the HT-29 and WiDr cell lines demonstrating an 8-fold increase. Significantly, the pattern of amplification observed did not correlate with that observed for c-*myc*, indicating that the c-*myc* gene is not part of the amplicon that involves the *WISP-1* locus.

We next examined whether the *WISP* genes were amplified in a panel of 25 primary human colon adenocarcinomas. The relative *WISP* gene copy number in each colon tumor DNA was compared with pooled normal DNA from 10 donors by quantitative PCR (Fig. 6). The copy number of *WISP-1* and *WISP-2* was significantly greater than one, approximately 2-fold for *WISP-1* in about 60% of the tumors and 2- to 4-fold for *WISP-2* in 92% of the tumors ($P < 0.001$ for each). The copy number for *WISP-3* was indistinguishable from one ($P = 0.166$). In addition, the copy number of *WISP-2* was significantly higher than that of *WISP-1* ($P < 0.001$).

The levels of *WISP* transcripts in RNA isolated from 19 adenocarcinomas and their matched normal mucosa were

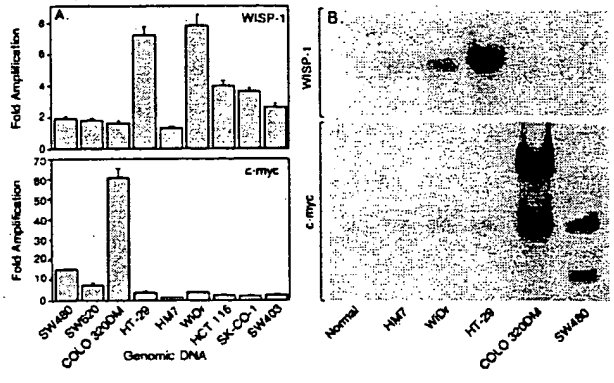


FIG. 5. Amplification of *WISP-1* genomic DNA in colon cancer cell lines. (*A*) Amplification in cell line DNA was determined by quantitative PCR. (*B*) Southern blots containing genomic DNA (10 μg) digested with *Eco*RI (*WISP-1*) or *Xba*I (c-*myc*) were hybridized with a 100-bp human *WISP-1* probe (amino acids 186–219) or a human c-*myc* probe (located at bp 1901–2000). The *WISP* and *myc* genes are detected in normal human genomic DNA after a longer film exposure.

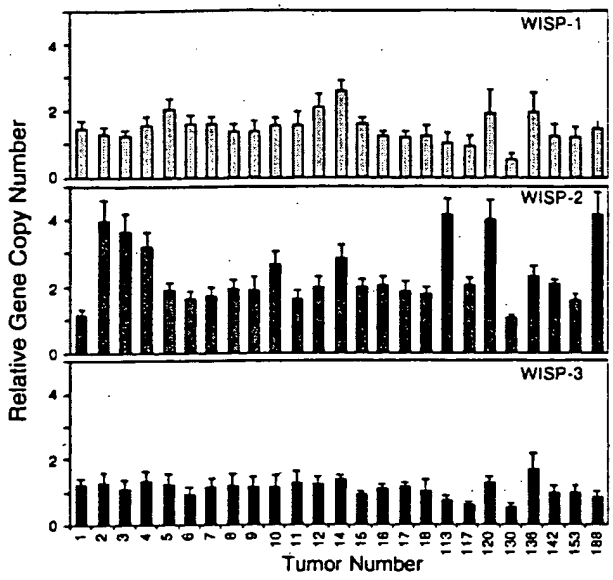


FIG. 6. Genomic amplification of *WISP* genes in human colon tumors. The relative gene copy number of the *WISP* genes in 25 adenocarcinomas was assayed by quantitative PCR, by comparing DNA from primary human tumors with pooled DNA from 10 healthy donors. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least three times.

assessed by quantitative PCR (Fig. 7). The level of *WISP-1* RNA present in tumor tissue varied but was significantly increased (2- to >25-fold) in 84% (16/19) of the human colon tumors examined compared with normal adjacent mucosa. Four of 19 tumors showed greater than 10-fold overexpression. In contrast, in 79% (15/19) of the tumors examined, *WISP-2* RNA expression was significantly lower in the tumor than the mucosa. Similar to *WISP-1*, *WISP-3* RNA was overexpressed in 63% (12/19) of the colon tumors compared with the normal

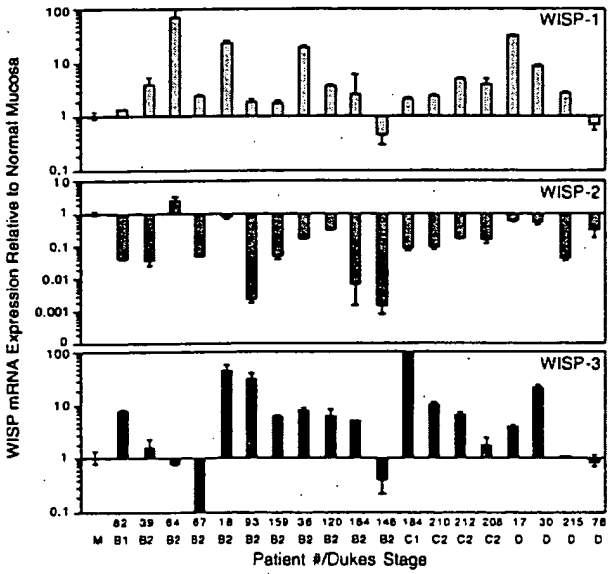


FIG. 7. *WISP* RNA expression in primary human colon tumors relative to expression in normal mucosa from the same patient. Expression of *WISP* mRNA in 19 adenocarcinomas was assayed by quantitative PCR. The Dukes stage of the tumor is listed under the sample number. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least twice.

mucosa. The amount of overexpression of *WISP-3* ranged from 4- to >40-fold.

DISCUSSION

One approach to understanding the molecular basis of cancer is to identify differences in gene expression between cancer cells and normal cells. Strategies based on assumptions that steady-state mRNA levels will differ between normal and malignant cells have been used to clone differentially expressed genes (31). We have used a PCR-based selection strategy, SSH, to identify genes selectively expressed in C57MG mouse mammary epithelial cells transformed by Wnt-1.

Three of the genes isolated, *WISP-1*, *WISP-2*, and *WISP-3*, are members of the CCN family of growth factors, which includes CTGF, Cyr61, and *nov*, a family not previously linked to Wnt signaling.

Two independent experimental systems demonstrated that *WISP* induction was associated with the expression of Wnt-1. The first was C57MG cells infected with a Wnt-1 retroviral vector or C57MG cells expressing Wnt-1 under the control of a tetracyline-repressible promoter, and the second was in Wnt-1 transgenic mice, where breast tissue expresses Wnt-1, whereas normal breast tissue does not. No *WISP* RNA expression was detected in mammary tumors induced by polyoma virus middle T antigen (data not shown). These data suggest a link between Wnt-1 and *WISP*s in that in these two situations, *WISP* induction was correlated with Wnt-1 expression.

It is not clear whether the *WISP*s are directly or indirectly induced by the downstream components of the Wnt-1 signaling pathway (i.e., β-catenin-TCF-1/Lef1). The increased levels of *WISP* RNA were measured in Wnt-1-transformed cells, hours or days after Wnt-1 transformation. Thus, *WISP* expression could result from Wnt-1 signaling directly through β-catenin transcription factor regulation or alternatively through Wnt-1 signaling turning on a transcription factor, which in turn regulates *WISP*s.

The *WISP*s define an additional subfamily of the CCN family of growth factors. One striking difference observed in the protein sequence of WISP-2 is the absence of a CT domain, which is present in CTGF, Cyr61, *nov*, WISP-1, and WISP-3. This domain is thought to be involved in receptor binding and dimerization. Growth factors, such as TGF-β, platelet-derived growth factor, and nerve growth factor, which contain a cystine knot motif exist as dimers (32). It is tempting to speculate that WISP-1 and WISP-3 may exist as dimers, whereas WISP-2 exists as a monomer. If the CT domain is also important for receptor binding, WISP-2 may bind its receptor through a different region of the molecule than the other CCN family members. No specific receptors have been identified for CTGF or *nov*. A recent report has shown that integrin $\alpha_v\beta_3$ serves as an adhesion receptor for Cyr61 (33).

The strong expression of *WISP-1* and *WISP-2* in cells lying within the fibrovascular tumor stroma in breast tumors from Wnt-1 transgenic animals is consistent with previous observations that transcripts for the related CTGF gene are primarily expressed in the fibrous stroma of mammary tumors (34). Epithelial cells are thought to control the proliferation of connective tissue stroma in mammary tumors by a cascade of growth factor signals similar to that controlling connective tissue formation during wound repair. It has been proposed that mammary tumor cells or inflammatory cells at the tumor interstitial interface secrete TGF-β1, which is the stimulus for stromal proliferation (34). TGF-β1 is secreted by a large percentage of malignant breast tumors and may be one of the growth factors that stimulates the production of CTGF and WISPs in the stroma.

It was of interest that *WISP-1* and *WISP-2* expression was observed in the stromal cells that surrounded the tumor cells

(epithelial cells) in the Wnt-1 transgenic mouse sections of breast tissue. This finding suggests that paracrine signaling could occur in which the stromal cells could supply WISP-1 and WISP-2 to regulate tumor cell growth on the WISP extracellular matrix. Stromal cell-derived factors in the extracellular matrix have been postulated to play a role in tumor cell migration and proliferation (35). The localization of *WISP-1* and *WISP-2* in the stromal cells of breast tumors supports this paracrine model.

An analysis of *WISP-1* gene amplification and expression in human colon tumors showed a correlation between DNA amplification and overexpression, whereas overexpression of *WISP-3* RNA was seen in the absence of DNA amplification. In contrast, *WISP-2* DNA was amplified in the colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with the expression in normal colonic mucosa from the same patient. The gene for human *WISP-2* was localized to chromosome 20q12–20q13, at a region frequently amplified and associated with poor prognosis in node negative breast cancer and many colon cancers, suggesting the existence of one or more oncogenes at this locus (36–38). Because the center of the 20q13 amplicon has not yet been identified, it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon.

A recent manuscript on *rCop-1*, the rat orthologue of *WISP-2*, describes the loss of expression of this gene after cell transformation, suggesting it may be a negative regulator of growth in cell lines (16). Although the mechanism by which *WISP-2* RNA expression is down-regulated during malignant transformation is unknown, the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. These results show that the *WISP* genes are aberrantly expressed in colon cancer and suggest that their altered expression may confer selective growth advantage to the tumor.

Members of the Wnt signaling pathway have been implicated in the pathogenesis of colon cancer, breast cancer, and melanoma, including the tumor suppressor gene adenomatous polyposis coli and β-catenin (39). Mutations in specific regions of either gene can cause the stabilization and accumulation of cytoplasmic β-catenin, which presumably contributes to human carcinogenesis through the activation of target genes such as the *WISP*s. Although the mechanism by which Wnt-1 transforms cells and induces tumorigenesis is unknown, the identification of *WISP*s as genes that may be regulated downstream of Wnt-1 in C57MG cells suggests they could be important mediators of Wnt-1 transformation. The amplification and altered expression patterns of the *WISP*s in human colon tumors may indicate an important role for these genes in tumor development.

We thank the DNA synthesis group for oligonucleotide synthesis, T. Baker for technical assistance, P. Dowd for radiation hybrid mapping, K. Willert and R. Nusse for the tet-repressible C57MG/Wnt-1 cells, V. Dixit for discussions, and D. Wood and A. Bruce for artwork.

1. Cadigan, K. M. & Nusse, R. (1997) *Genes Dev.* **11**, 3286–3305.
2. Dale, T. C. (1998) *Biochem. J.* **329**, 209–223.
3. Nusse, R. & Varmus, H. E. (1982) *Cell* **31**, 99–109.
4. van Ooyen, A. & Nusse, R. (1984) *Cell* **39**, 233–240.
5. Tsukamoto, A. S., Grosschedl, R., Guzman, R. C., Parslow, T. & Varmus, H. E. (1988) *Cell* **55**, 619–625.
6. Brown, J. D. & Moon, R. T. (1998) *Curr. Opin. Cell. Biol.* **10**, 182–187.
7. Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) *Cell* **86**, 391–399.
8. Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) *Mol. Cell. Biol.* **18**, 1248–1256.
9. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 3046–3050.
10. He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* **281**, 1509–1512.
11. Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. & Siebert, P. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 6025–6030.
12. Brown, A. M., Wildin, R. S., Prendergast, T. J. & Varmus, H. E. (1986) *Cell* **46**, 1001–1009.
13. Wong, G. T., Gavin, B. J. & McMahon, A. P. (1994) *Mol. Cell. Biol.* **14**, 6278–6286.
14. Shimizu, H., Julius, M. A., Giarre, M., Zheng, Z., Brown, A. M. & Kitajewski, J. (1997) *Cell Growth Differ.* **8**, 1349–1358.
15. Hashimoto, Y., Shindo-Okada, N., Tani, M., Nagamachi, Y., Takeuchi, K., Shiroishi, T., Toma, H. & Yokota, J. (1998) *J. Exp. Med.* **187**, 289–296.
16. Zhang, R., Averboukh, L., Zhu, W., Zhang, H., Jo, H., Dempsey, P. J., Coffey, R. J., Pardee, A. B. & Liang, P. (1998) *Mol. Cell. Biol.* **18**, 6131–6141.
17. Grotendorst, G. R. (1997) *Cytokine Growth Factor Rev.* **8**, 171–179.
18. Kireeva, M. L., Mo, F. E., Yang, G. P. & Lau, L. F. (1996) *Mol. Cell. Biol.* **16**, 1326–1334.
19. Babic, A. M., Kireeva, M. L., Kolesnikova, T. V. & Lau, L. F. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 6355–6360.
20. Martinerie, C., Huff, V., Joubert, I., Badzioch, M., Saunders, G., Strong, L. & Perbal, B. (1994) *Oncogene* **9**, 2729–2732.
21. Bork, P. (1993) *FEBS Lett.* **327**, 125–130.
22. Kim, H. S., Nagalla, S. R., Oh, Y., Wilson, E., Roberts, C. T., Jr. & Rosenfeld, R. G. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 12981–12986.
23. Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992) *Mol. Cell. Biol.* **12**, 10–21.
24. Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. & Sadler, J. E. (1989) *J. Biol. Chem.* **264**, 19514–19527.
25. Holt, G. D., Pangburn, M. K. & Ginsburg, V. (1990) *J. Biol. Chem.* **265**, 2852–2855.
26. Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991) *J. Cell. Biol.* **113**, 195–205.
27. Martinerie, C., Viegas-Pequignot, E., Guenard, I., Dutrillaux, B., Nguyen, V. C., Bernheim, A. & Perbal, B. (1992) *Oncogene* **7**, 2529–2534.
28. Takahashi, E., Hori, T., O'Connell, P., Leppert, M. & White, R. (1991) *Cytogenet. Cell. Genet.* **57**, 109–111.
29. Meese, E., Meltzer, P. S., Witkowski, C. M. & Trent, J. M. (1989) *Genes Chromosomes Cancer* **1**, 88–94.
30. Garte, S. J. (1993) *Crit. Rev. Oncog.* **4**, 435–449.
31. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276**, 1268–1272.
32. Sun, P. D. & Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* **24**, 269–291.
33. Kireeva, M. L., Lam, S. C. T. & Lau, L. F. (1998) *J. Biol. Chem.* **273**, 3090–3096.
34. Frazier, K. S. & Grotendorst, G. R. (1997) *Int. J. Biochem. Cell. Biol.* **29**, 153–161.
35. Wernert, N. (1997) *Virchows Arch.* **430**, 433–443.
36. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., *et al.* (1994) *Cancer Res.* **54**, 4257–4260.
37. Brinkmann, U., Gallo, M., Polymeropoulos, M. H. & Pastan, I. (1996) *Genome Res.* **6**, 187–194.
38. Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., *et al.* (1998) *EMBO J.* **17**, 3052–3065.
39. Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **275**, 1787–1790.
40. Lu, L. H. & Gillett, N. (1994) *Cell Vision* **1**, 169–176.

methods. Peptides AENK or AEQK were dissolved in water, made isotonic with NaCl and diluted into RPMI growth medium. T-cell-proliferation assays were done essentially as described[20,21]. Briefly, after antigen pulsing (30 μg ml^{-1} TTCF) with tetrapeptides (1–2 mg ml^{-1}), PBMCs or EBV-B cells were washed in PBS and fixed for 45 s in 0.05% glutaraldehyde. Glycine was added to a final concentration of 0.1M and the cells were washed five times in RPMI 1640 medium containing 1% FCS before co-culture with T-cell clones in round-bottom 96-well microtitre plates. After 48 h, the cultures were pulsed with 1 μCi of ^{3}H-thymidine and harvested for scintillation counting 16 h later. Predigestion of native TTCF was done by incubating 200 μg TTCF with 0.25 μg pig kidney legumain in 500 μl 50 mM citrate buffer, pH 5.5, for 1 h at 37 °C.

Glycopeptide digestions. The peptides HIDNEEDI, HIDN(*N*-glucosamine) EEDI and HIDNESDI, which are based on the TTCF sequence, and QQQHLFGSNVTDCSGNFCLFR(KKK), which is based on human transferrin, were obtained by custom synthesis. The three C-terminal lysine residues were added to the natural sequence to aid solubility. The transferrin glycopeptide QQQHLFGSNVTDCSGNFCLFR was prepared by tryptic (Promega) digestion of 5 mg reduced, carboxy-methylated human transferrin followed by concanavalin A chromatography[11]. Glycopeptides corresponding to residues 622–642 and 421–452 were isolated by reverse-phase HPLC and identified by mass spectrometry and N-terminal sequencing. The lyophilized transferrin-derived peptides were redissolved in 50 mM sodium acetate, pH 5.5, 10 mM dithiothreitol, 20% methanol. Digestions were performed for 3 h at 30 °C with 5–50 mU ml^{-1} pig kidney legumain or B-cell AEP. Products were analysed by HPLC or MALDI-TOF mass spectrometry using a matrix of 10 mg ml^{-1} α-cyanocinnamic acid in 50% acetonitrile/0.1% TFA and a PerSeptive Biosystems Elite STR mass spectrometer set to linear or reflector mode. Internal standardization was obtained with a matrix ion of 568.13 mass units.

Received 29 September; accepted 3 November 1998.

1. Chen, J. M. *et al.* Cloning, isolation, and characterisation of mammalian legumain, an asparaginyl endopeptidase. *J. Biol. Chem.* 272, 8090–8098 (1997).
2. Kembhavi, A. A., Buttle, D. J., Knight, C. G. & Barrett, A. J. The two cysteine endopeptidases of legume seeds: purification and characterization by use of specific fluorometric assays. *Arch. Biochem. Biophys.* 303, 208–213 (1993).
3. Dalton, J. P., Hola Jamriska, L. & Bridley, P. J. Asparaginyl endopeptidase activity in adult *Schistosoma mansoni*. *Parasitology* 111, 575–580 (1995).
4. Bennett, K. *et al.* Antigen processing for presentation by class II major histocompatibility complex requires cleavage by cathespin E. *Eur. J. Immunol.* 22, 1519–1524 (1992).
5. Riese, R. J. *et al.* Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading. *Immunity* 4, 357–366 (1996).
6. Rodriguez, G. M. & Diment, S. Role of cathepsin D in antigen presentation of ovalbumin. *J. Immunol.* 149, 2894–2898 (1992).
7. Hewitt, E. W. *et al.* Natural processing sites for human cathepsin E and cathepsin D in tetanus toxin: implications for T cell epitope generation. *J. Immunol.* 159, 4693–4699 (1997).
8. Watts, C. Capture and processing of exogenous antigens for presentation on MHC molecules. *Annu. Rev. Immunol.* 15, 821–850 (1997).
9. Chapman, H. A. Endosomal proteases and MHC class II function. *Curr. Opin. Immunol.* 10, 93–102 (1998).
10. Fineschi, B. & Miller, J. Endosomal proteases and antigen processing. *Trends Biochem. Sci.* 22, 377–382 (1997).
11. Lu, J. & van Halbeek, H. Complete ^{1}H and ^{13}C resonance assignments of a 21-amino acid glycopeptide prepared from human serum transferrin. *Carbohydr. Res.* 296, 1–21 (1996).
12. Fearon, D. T. & Locksley, R. M. The instructive role of innate immunity in the acquired immune response. *Science* 272, 50–54 (1996).
13. Medzhitov, R. & Janeway, C. A. J. Innate immunity: the virtues of a nonclonal system of recognition. *Cell* 91, 295–298 (1997).
14. Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* 393, 705–711 (1998).
15. Botarelli, P. *et al.* N-glycosylation of HIV gp120 may constrain recognition by T lymphocytes. *J. Immunol.* 147, 3128–3132 (1991).
16. Davidson, H. W., West, M. A. & Watts, C. Endocytosis, intracellular trafficking, and processing of membrane IgG and monovalent antigen/membrane IgG complexes in B lymphocytes. *J. Immunol.* 144, 4101–4109 (1990).
17. Barrett, A. J. & Kirschke, H. Cathepsin B, cathepsin H and cathepsin L. *Methods Enzymol.* 80, 535–559 (1981).
18. Makoff, A. J., Ballantine, S. P., Smallwood, A. E. & Fairweather, N. F. Expression of tetanus toxin fragment C in *E. coli*: its purification and potential use as a vaccine. *Biotechnology* 7, 1043–1046 (1989).
19. Lane, D. P. & Harlow, E. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 1988).
20. Lanzavecchia, A. Antigen-specific interaction between T and B cells. *Nature* 314, 537–539 (1985).
21. Pond, L. & Watts, C. Characterization of transport of newly assembled, T cell-stimulatory MHC class II-peptide complexes from MHC class II compartments to the cell surface. *J. Immunol.* 159, 543–553 (1997).

Acknowledgements. We thank M. Ferguson for helpful discussions and advice; E. Smythe and L. Grayson for advice and technical assistance; B. Spruce, A. Knight and the BTS (Ninewells Hospital) for help with blood monocyte preparation; and our colleagues for many helpful comments on the manuscript. This work was supported by the Wellcome Trust and by an EMBO Long-term fellowship to B. M.

Correspondence and requests for materials should be addressed to C.W. (e-mail: c.watts@dundee.ac.uk).

Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer

Robert M. Pitti*†, Scot A. Marsters*†, David A. Lawrence*†, Margaret Roy*, Frank C. Kischkel*, Patrick Dowd*, Arthur Huang*, Christopher J. Donahue*, Steven W. Sherwood*, Daryl T. Baldwin*, Paul J. Godowski*, William I. Wood*, Austin L. Gurney*, Kenneth J. Hillan*, Robert L. Cohen*, Audrey D. Goddard*, David Botstein‡ & Avi Ashkenazi*

* *Departments of Molecular Oncology, Molecular Biology, and Immunology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA*
‡ *Department of Genetics, Stanford University, Stanford, California 94305, USA*
† *These authors contributed equally to this work*

Fas ligand (FasL) is produced by activated T cells and natural killer cells and it induces apoptosis (programmed cell death) in target cells through the death receptor Fas/Apo1/CD95 (ref. 1). One important role of FasL and Fas is to mediate immune-cytotoxic killing of cells that are potentially harmful to the organism, such as virus-infected or tumour cells[1]. Here we report the discovery of a soluble decoy receptor, termed decoy receptor 3 (DcR3), that binds to FasL and inhibits FasL-induced apoptosis. The DcR3 gene was amplified in about half of 35 primary lung and colon tumours studied, and DcR3 messenger RNA was expressed in malignant tissue. Thus, certain tumours may escape FasL-dependent immune-cytotoxic attack by expressing a decoy receptor that blocks FasL.

By searching expressed sequence tag (EST) databases, we identified a set of related ESTs that showed homology to the tumour necrosis factor (TNF) receptor (TNFR) gene superfamily[2]. Using the overlapping sequence, we isolated a previously unknown full-length complementary DNA from human fetal lung. We named the protein encoded by this cDNA decoy receptor 3 (DcR3). The cDNA encodes a 300-amino-acid polypeptide that resembles members of the TNFR family (Fig. 1a): the amino terminus contains a leader sequence, which is followed by four tandem cysteine-rich domains (CRDs). Like one other TNFR homologue, osteoprotegerin (OPG)[3], DcR3 lacks an apparent transmembrane sequence, which indicates that it may be a secreted, rather than a membrane-asscociated, molecule. We expressed a recombinant, histidine-tagged form of DcR3 in mammalian cells; DcR3 was secreted into the cell culture medium, and migrated on polyacrylamide gels as a protein of relative molecular mass 35,000 (data not shown). DcR3 shares sequence identity in particular with OPG (31%) and TNFR2 (29%), and has relatively less homology with Fas (17%). All of the cysteines in the four CRDs of DcR3 and OPG are conserved; however, the carboxy-terminal portion of DcR3 is 101 residues shorter.

We analysed expression of DcR3 mRNA in human tissues by northern blotting (Fig. 1b). We detected a predominant 1.2-kilobase transcript in fetal lung, brain, and liver, and in adult spleen, colon and lung. In addition, we observed relatively high DcR3 mRNA expression in the human colon carcinoma cell line SW480.

To investigate potential ligand interactions of DcR3, we generated a recombinant, Fc-tagged DcR3 protein. We tested binding of DcR3–Fc to human 293 cells transfected with individual TNF-family ligands, which are expressed as type 2 transmembrane proteins (these transmembrane proteins have their N termini in the cytosol). DcR3–Fc showed a significant increase in binding to cells transfected with FasL[4] (Fig. 2a), but not to cells transfected with TNF[5], Apo2L/TRAIL[6,7], Apo3L/TWEAK[8,9], or OPGL/TRANCE/

RANKL[10-12] (data not shown). DcR3–Fc immunoprecipitated shed FasL from FasL-transfected 293 cells (Fig. 2b) and purified soluble FasL (Fig. 2c), as did the Fc-tagged ectodomain of Fas but not TNFR1. Gel-filtration chromatography showed that DcR3–Fc and soluble FasL formed a stable complex (Fig. 2d). Equilibrium analysis indicated that DcR3–Fc and Fas–Fc bound to soluble FasL with a comparable affinity ($K_d = 0.8 \pm 0.2$ and 1.1 ± 0.1 nM, respectively; Fig. 2e), and that DcR3–Fc could block nearly all of the binding of soluble FasL to Fas–Fc (Fig. 2e, inset). Thus, DcR3 competes with Fas for binding to FasL.

To determine whether binding of DcR3 inhibits FasL activity, we tested the effect of DcR3–Fc on apoptosis induction by soluble FasL in Jurkat T leukaemia cells, which express Fas (Fig. 3a). DcR3–Fc and Fas–Fc blocked soluble-FasL-induced apoptosis in a similar dose-dependent manner, with half-maximal inhibition at $\sim 0.1\ \mu g\ ml^{-1}$. Time-course analysis showed that the inhibition did not merely delay cell death, but rather persisted for at least 24 hours (Fig. 3b). We also tested the effect of DcR3–Fc on activation-induced cell death (AICD) of mature T lymphocytes, a FasL-dependent process[1]. Consistent with previous results[13], activation of interleukin-2-stimulated CD4-positive T cells with anti-CD3 antibody increased the level of apoptosis twofold, and Fas–Fc blocked this effect substantially (Fig. 3c); DcR3–Fc blocked the induction of apoptosis to a similar extent. Thus, DcR3 binding blocks apoptosis induction by FasL.

FasL-induced apoptosis is important in elimination of virus-infected cells and cancer cells by natural killer cells and cytotoxic T lymphocytes; an alternative mechanism involves perforin and granzymes[1,14-16]. Peripheral blood natural killer cells triggered marked cell death in Jurkat T leukaemia cells (Fig. 3d); DcR3–Fc and Fas–Fc each reduced killing of target cells from ~65% to ~30%, with half-maximal inhibition at $\sim 1\ \mu g\ ml^{-1}$; the residual killing was probably mediated by the perforin/granzyme pathway. Thus, DcR3 binding blocks FasL-dependent natural killer cell activity. Higher DcR3–Fc and Fas–Fc concentrations were required to block natural killer cell activity compared with those required to block soluble FasL activity, which is consistent with the greater potency of membrane-associated FasL compared with soluble FasL[17].

Given the role of immune-cytotoxic cells in elimination of tumour cells and the fact that DcR3 can act as an inhibitor of FasL, we proposed that DcR3 expression might contribute to the ability of some tumours to escape immune-cytotoxic attack. As genomic amplification frequently contributes to tumorigenesis, we investigated whether the DcR3 gene is amplified in cancer. We analysed DcR3 gene-copy number by quantitative polymerase chain

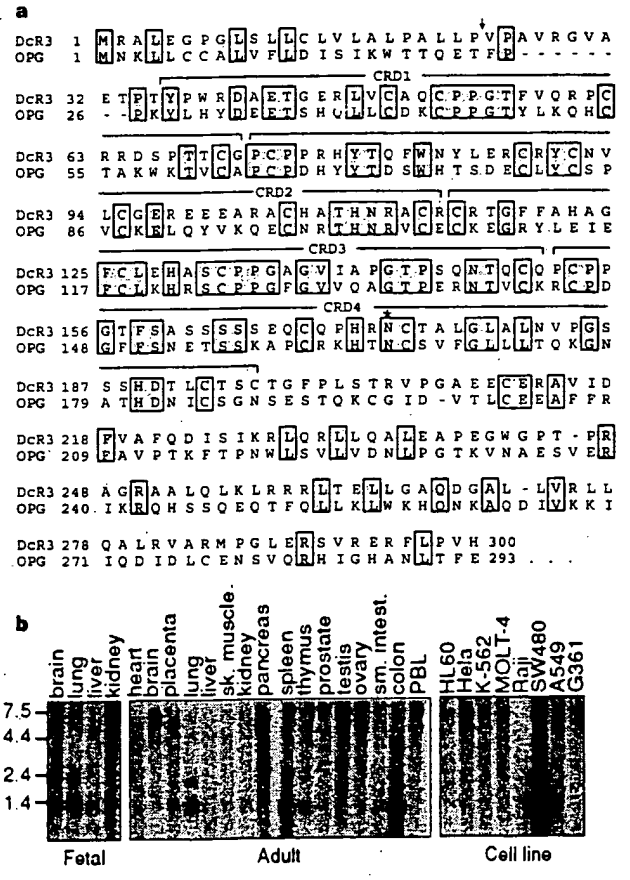


Figure 1 Primary structure and expression of human DcR3. **a**, Alignment of the amino-acid sequences of DcR3 and of osteoprotegerin (OPG); the C-terminal 101 residues of OPG are not shown. The putative signal cleavage site (arrow), the cysteine-rich domains (CRD 1-4), and the *N*-linked glycosylation site (asterisk) are shown. **b**, Expression of DcR3 mRNA. Northern hybridization analysis was done using the DcR3 cDNA as a probe and blots of poly(A)$^+$ RNA (Clontech) from human fetal and adult tissues or cancer cell lines. PBL, peripheral blood lymphocyte.

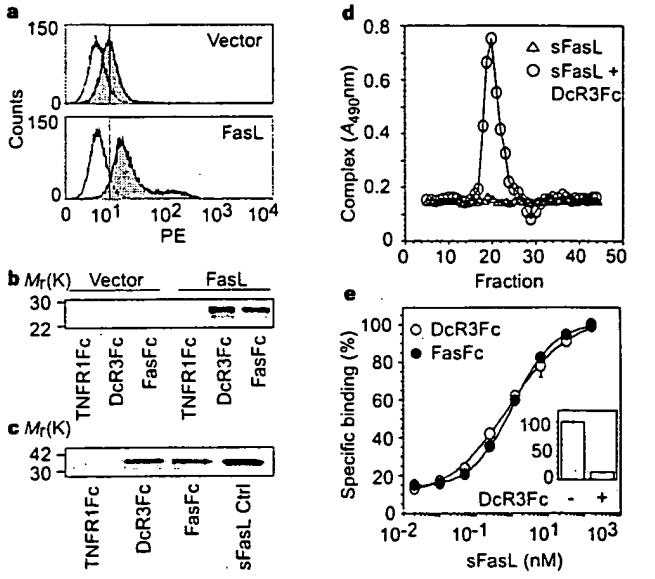


Figure 2 Interaction of DcR3 with FasL. **a**, 293 cells were transfected with pRK5 vector (top) or with pRK5 encoding full-length FasL (bottom), incubated with DcR3-Fc (solid line, shaded area), TNFR1-Fc (dotted line) or buffer control (dashed line) (the dashed and dotted lines overlap), and analysed for binding by FACS. Statistical analysis showed a significant difference ($P < 0.001$) between the binding of DcR3-Fc to cells transfected with FasL or pRK5. PE, phycoerythrin-labelled cells. **b**, 293 cells were transfected as in **a** and metabolically labelled, and cell supernatants were immunoprecipitated with Fc-tagged TNFR1, DcR3 or Fas. **c**, Purified soluble FasL (sFasL) was immunoprecipitated with TNFR1-Fc, DcR3-Fc or Fas-Fc and visualized by immunoblot with anti-FasL antibody. sFasL was loaded directly for comparison in the right-hand lane. **d**, Flag-tagged sFasL was incubated with DcR3-Fc or with buffer and resolved by gel filtration; column fractions were analysed in an assay that detects complexes containing DcR3-Fc and sFasL-Flag. **e**, Equilibrium binding of DcR3-Fc or Fas-Fc to sFasL-Flag. Inset, competition of DcR3-Fc with Fas-Fc for binding to sFasL-Flag.

reaction (PCR)[18] in genomic DNA from 35 primary lung and colon tumours, relative to pooled genomic DNA from peripheral blood leukocytes (PBLs) of 10 healthy donors. Eight of 18 lung tumours and 9 of 17 colon tumours showed DcR3 gene amplification, ranging from 2- to 18-fold (Fig. 4a, b). To confirm this result, we analysed the colon tumour DNAs with three more, independent sets of DcR3-based PCR primers and probes; we observed nearly the same amplification (data not shown).

We then analysed DcR3 mRNA expression in primary tumour tissue sections by *in situ* hybridization. We detected DcR3 expression in 6 out of 15 lung tumours, 2 out of 2 colon tumours, 2 out of 5 breast tumours, and 1 out of 1 gastric tumour (data not shown). A section through a squamous-cell carcinoma of the lung is shown in Fig. 4c. DcR3 mRNA was localized to infiltrating malignant epithelium, but was essentially absent from adjacent stroma, indicating tumour-specific expression. Although the individual tumour specimens that we analysed for mRNA expression and gene amplification were different, the *in situ* hybridization results are consistent with the finding that the DcR3 gene is amplified frequently in tumours. SW480 colon carcinoma cells, which showed abundant DcR3 mRNA expression (Fig. 1b), also had marked DcR3 gene amplification, as shown by quantitative PCR (fourfold) and by Southern blot hybridization (fivefold) (data not shown).

If DcR3 amplification in cancer is functionally relevant, then DcR3 should be amplified more than neighbouring genomic regions that are not important for tumour survival. To test this, we mapped the human DcR3 gene by radiation-hybrid analysis; DcR3 showed linkage to marker AFM218xe7 (T160), which maps to chromosome position 20q13. Next, we isolated from a bacterial artificial chromosome (BAC) library a human genomic clone that carries DcR3, and sequenced the ends of the clone's insert. We then determined, from the nine colon tumours that showed twofold or greater amplification of DcR3, the copy number of the DcR3-flanking sequences (reverse and forward) from the BAC, and of seven genomic markers that span chromosome 20 (Fig. 4d). The DcR3-linked reverse marker showed an average amplification of roughly threefold, slightly less than the approximately fourfold amplification of DcR3; the other markers showed little or no amplification. These data indicate that DcR3 may be at the 'epicentre' of a distal chromosome 20 region that is amplified in colon cancer, consistent with the possibility that DcR3 amplification promotes tumour survival.

Our results show that DcR3 binds specifically to FasL and inhibits FasL activity. We did not detect DcR3 binding to several other TNF-ligand-family members; however, this does not rule out the possibility that DcR3 interacts with other ligands, as do some other TNFR family members, including OPG[2,19].

FasL is important in regulating the immune response; however, little is known about how FasL function is controlled. One mechanism involves the molecule cFLIP, which modulates apoptosis signalling downstream of Fas[20]. A second mechanism involves proteolytic shedding of FasL from the cell surface[17]. DcR3 competes with Fas for

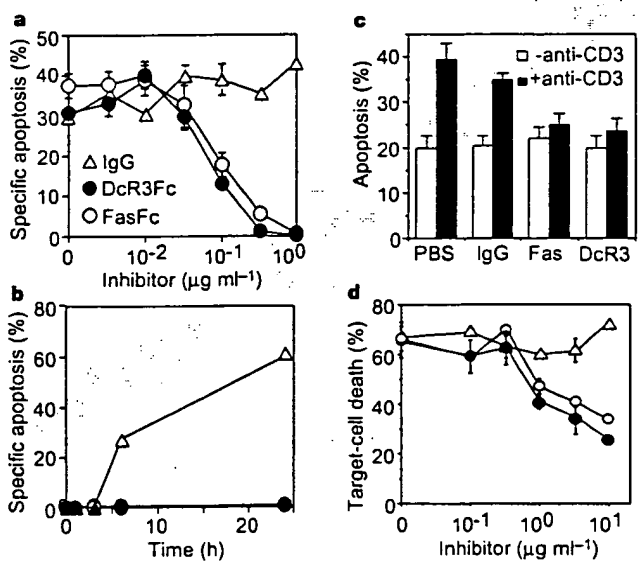


Figure 3 Inhibition of FasL activity by DcR3. **a**, Human Jurkat T leukaemia cells were incubated with Flag-tagged soluble FasL (sFasL; 5 ng ml^{-1}) oligomerized with anti-Flag antibody (0.1 μg ml^{-1}) in the presence of the proposed inhibitors DcR3-Fc, Fas-Fc or human IgG1 and assayed for apoptosis (mean ± s.e.m. of triplicates). **b**, Jurkat cells were incubated with sFasL-Flag plus anti-Flag antibody as in **a**, in presence of 1 μg ml^{-1} DcR3-Fc (filled circles), Fas-Fc (open circles) or human IgG1 (triangles), and apoptosis was determined at the indicated time points. **c**, Peripheral blood T cells were stimulated with PHA and interleukin-2, followed by control (white bars) or anti-CD3 antibody (filled bars), together with phosphate-buffered saline (PBS), human IgG1, Fas-Fc, or DcR3-Fc (10 μg ml^{-1}). After 16 h, apoptosis of CD4$^+$ cells was determined (mean ± s.e.m. of results from five donors). **d**, Peripheral blood natural killer cells were incubated with ^{51}Cr-labelled Jurkat cells in the presence of DcR3-Fc (filled circles), Fas-Fc (open circles) or human IgG1 (triangles), and target-cell death was determined by release of ^{51}Cr (mean ± s.d. for two donors, each in triplicate).

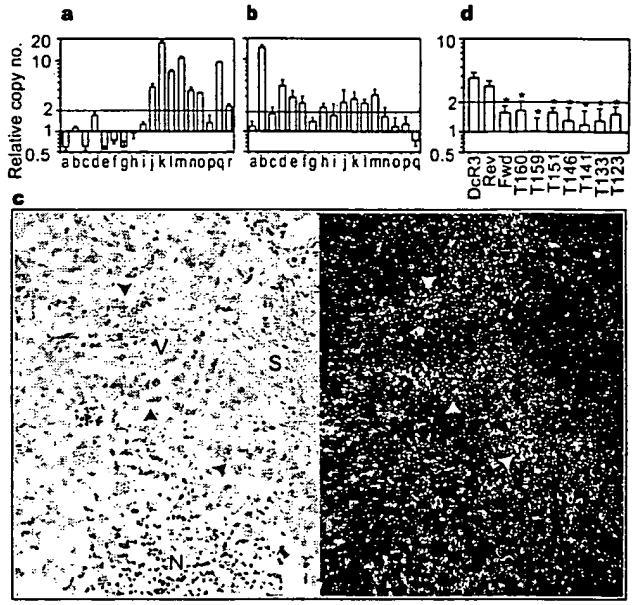


Figure 4 Genomic amplification of DcR3 in tumours. **a**, Lung cancers, comprising eight adenocarcinomas (c, d, f, g, h, j, k, r), seven squamous-cell carcinomas (a, e, m, n, o, p, q), one non-small-cell carcinoma (b), one small-cell carcinoma (i), and one bronchial adenocarcinoma (l). The data are means ± s.d. of 2 experiments done in duplicate. **b**, Colon tumours, comprising 17 adenocarcinomas. Data are means ± s.e.m. of five experiments done in duplicate. **c**, *In situ* hybridization analysis of DcR3 mRNA expression in a squamous-cell carcinoma of the lung. A representative bright-field image (left) and the corresponding dark-field image (right) show DcR3 mRNA over infiltrating malignant epithelium (arrowheads). Adjacent non-malignant stroma (S), blood vessel (V) and necrotic tumour tissue (N) are also shown. **d**, Average amplification of DcR3 compared with amplification of neighbouring genomic regions (reverse and forward, Rev and Fwd), the DcR3-linked marker T160, and other chromosome-20 markers, in the nine colon tumours showing DcR3 amplification of twofold or more (**b**). Data are from two experiments done in duplicate. Asterisk indicates $P < 0.01$ for a Student's *t*-test comparing each marker with DcR3.

FasL binding; hence, it may represent a third mechanism of extracellular regulation of FasL activity. A decoy receptor that modulates the function of the cytokine interleukin-1 has been described[21]. In addition, two decoy receptors that belong to the TNFR family, DcR1 and DcR2, regulate the FasL-related apoptosis-inducing molecule Apo2L[22]. Unlike DcR1 and DcR2, which are membrane-associated proteins, DcR3 is directly secreted into the extracellular space. One other secreted TNFR-family member is OPG[3], which shares greater sequence homology with DcR3 (31%) than do DcR1 (17%) or DcR2 (19%); OPG functions as a third decoy for Apo2L[19]. Thus, DcR3 and OPG define a new subset of TNFR-family members that function as secreted decoys to modulate ligands that induce apoptosis. Pox viruses produce soluble TNFR homologues that neutralize specific TNF-family ligands, thereby modulating the antiviral immune response[2]. Our results indicate that a similar mechanism, namely, production of a soluble decoy receptor for FasL, may contribute to immune evasion by certain tumours. □

Methods

Isolation of DcR3 cDNA. Several overlapping ESTs in GenBank (accession numbers AA025672, AA025673 and W67560) and in Lifeseq™ (Incyte Pharmaceuticals; accession numbers 1339238, 1533571, 1533650, 1542861, 1789372 and 2207027) showed similarity to members of the TNFR family. We screened human cDNA libraries by PCR with primers based on the region of EST consensus; fetal lung was positive for a product of the expected size. By hybridization to a PCR-generated probe based on the ESTs, one positive clone (DNA30942) was identified. When searching for potential alternatively spliced forms of DcR3 that might encode a transmembrane protein, we isolated 50 more clones; the coding regions of these clones were identical in size to that of the initial clone (data not shown).

Fc-fusion proteins (immunoadhesins). The entire DcR3 sequence, or the ectodomain of Fas or TNFR1, was fused to the hinge and Fc region of human IgG1, expressed in insect SF9 cells or in human 293 cells, and purified as described[23].

Fluorescence-activated cell sorting (FACS) analysis. We transfected 293 cells using calcium phosphate or Effectene (Qiagen) with pRK5 vector or pRK5 encoding full-length human FasL[4] (2 μg), together with pRK5 encoding CrmA (2 μg) to prevent cell death. After 16 h, the cells were incubated with biotinylated DcR3–Fc or TNFR1–Fc and then with phycoerythrin-conjugated streptavidin (GibcoBRL), and were assayed by FACS. The data were analysed by Kolmogorov–Smirnov statistical analysis. There was some detectable staining of vector-transfected cells by DcR3–Fc; as these cells express little FasL (data not shown), it is possible that DcR3 recognized some other factor that is expressed constitutively on 293 cells.

Immunoprecipitation. Human 293 cells were transfected as above, and metabolically labelled with [^{35}S]cysteine and [^{35}S]methionine (0.5 mCi; Amersham). After 16 h of culture in the presence of z-VAD-fmk (10 μM), the medium was immunoprecipitated with DcR3–Fc, Fas–Fc or TNFR1–Fc (5 μg), followed by protein A–Sepharose (Repligen). The precipitates were resolved by SDS–PAGE and visualized on a phosphorimager (Fuji BAS2000). Alternatively, purified, Flag-tagged soluble FasL (1 μg) (Alexis) was incubated with each Fc-fusion protein (1 μg), precipitated with protein A–Sepharose, resolved by SDS–PAGE and visualized by immunoblotting with rabbit anti-FasL antibody (Oncogene Research).

Analysis of complex formation. Flag-tagged soluble FasL (25 μg) was incubated with buffer or with DcR3–Fc (40 μg) for 1.5 h at 24 °C. The reaction was loaded onto a Superdex 200 HR 10/30 column (Pharmacia) and developed with PBS; 0.6-ml fractions were collected. The presence of DcR3–Fc–FasL complex in each fraction was analysed by placing 100 μl aliquots into microtitre wells precoated with anti-human IgG (Boehringer) to capture DcR3–Fc, followed by detection with biotinylated anti-Flag antibody Bio M2 (Kodak) and streptavidin–horseradish peroxidase (Amersham). Calibration of the column indicated an apparent relative molecular mass of the complex of 420K (data not shown), which is consistent with a stoichiometry of two DcR3–Fc homodimers to two soluble FasL homotrimers.

Equilibrium binding analysis. Microtitre wells were coated with anti-human IgG, blocked with 2% BSA in PBS. DcR3–Fc or Fas–Fc was added, followed by serially diluted Flag-tagged soluble FasL. Bound ligand was detected with anti-Flag antibody as above. In the competition assay, Fas–Fc was immobilized as above, and the wells were blocked with excess IgG1 before addition of Flag-tagged soluble FasL plus DcR3–Fc.

T-cell AICD. CD3$^+$ lymphocytes were isolated from peripheral blood of individual donors using anti-CD3 magnetic beads (Miltenyi Biotech), stimulated with phytohaemagglutinin (PHA; 2 μg ml^{-1}) for 24 h, and cultured in the presence of interleukin-2 (100 U ml^{-1}) for 5 days. The cells were plated in wells coated with anti-CD3 antibody (Pharmingen) and analysed for apoptosis 16 h later by FACS analysis of annexin-V-binding of CD4$^+$ cells[24].

Natural killer cell activity. Natural killer cells were isolated from peripheral blood of individual donors using anti-CD56 magnetic beads (Miltenyi Biotech), and incubated for 16 h with ^{51}Cr-loaded Jurkat cells at an effector-to-target ratio of 1:1 in the presence of DcR3–Fc, Fas–Fc or human IgG1. Target-cell death was determined by release of ^{51}Cr in effector–target co-cultures relative to release of ^{51}Cr by detergent lysis of equal numbers of Jurkat cells.

Gene-amplification analysis. Surgical specimens were provided by J. Kern (lung tumours) and P. Quirke (colon tumours). Genomic DNA was extracted (Qiagen) and the concentration was determined using Hoechst dye 33258 intercalation fluorometry. Amplification was determined by quantitative PCR[18] using a TaqMan instrument (ABI). The method was validated by comparison of PCR and Southern hybridization data for the Myc and HER-2 oncogenes (data not shown). Gene-specific primers and fluorogenic probes were designed on the basis of the sequence of DcR3 or of nearby regions identified on a BAC carrying the human DcR3 gene; alternatively, primers and probes were based on Stanford Human Genome Center marker AFM218xe7 (T160), which is linked to DcR3 (likelihood score = 5.4), SHGC-36268 (T159), the nearest available marker which maps to ~500 kilobases from T160, and five extra markers that span chromosome 20. The DcR3-specific primer sequences were 5′-CTTCTTCGCGCACGCTG-3′ and 5′-ATCACGCCGGCACCAG-3′ and the fluorogenic probe sequence was 5′-(FAM-ACACGATGCGTGCTCCAAGCAG AAp-(TAMARA), where FAM is 5′-fluorescein phosphoramidite. Relative gene-copy numbers were derived using the formula $2^{(\Delta CT)}$, where ΔCT is the difference in amplification cycles required to detect DcR3 in peripheral blood lymphocyte DNA compared to test DNA.

Received 24 September; accepted 6 November 1998.

1. Nagata, S. Apoptosis by death factor. *Cell* 88, 355–365 (1997).
2. Smith, C. A., Farrah, T. & Goodwin, R. G. The TNF receptor superfamily of cellular and viral proteins: activation, costimulation, and death. *Cell* 76, 959–962 (1994).
3. Simonet, W. S. *et al.* Osteoprotegerin: a novel secreted protein involved in the regulation of bone density. *Cell* 89, 309–319 (1997).
4. Suda, T., Takahashi, T., Golstein, P. & Nagata, S. Molecular cloning and expression of Fas ligand, a novel member of the TNF family. *Cell* 75, 1169–1178 (1993).
5. Pennica, D. *et al.* Human tumour necrosis factor: precursor structure, expression and homology to lymphotoxin. *Nature* 312, 724–729 (1984).
6. Pitti, R. M. *et al.* Induction of apoptosis by Apo-2 ligand, a new member of the tumor necrosis factor receptor family. *J. Biol. Chem.* 271, 12687–12690 (1996).
7. Wiley, S. R. *et al.* Identification and characterization of a new member of the TNF family that induces apoptosis. *Immunity* 3, 673–682 (1995).
8. Marsters, S. A. *et al.* Identification of a ligand for the death-domain-containing receptor Apo3. *Curr. Biol.* 8, 525–528 (1998).
9. Chicheportiche, Y. *et al.* TWEAK, a new secreted ligand in the TNF family that weakly induces apoptosis. *J. Biol. Chem.* 272, 32401–32410 (1997).
10. Wong, B. R. *et al.* TRANCE is a novel ligand of the TNFR family that activates c-Jun-N-terminal kinase in T cells. *J. Biol. Chem.* 272, 25190–25194 (1997).
11. Anderson, D. M. *et al.* A homolog of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function. *Nature* 390, 175–179 (1997).
12. Lacey, D. L. *et al.* Osteoprotegerin ligand is a cytokine that regulates osteoclast differentiation and activation. *Cell* 93, 165–176 (1998).
13. Dhein, J., Walczak, H., Baumler, C., Debatin, K. M. & Krammer, P. H. Autocrine T-cell suicide mediated by Apo1/(Fas/CD95). *Nature* 373, 438–441 (1995).
14. Arase, H., Arase, N. & Saito, T. Fas-mediated cytotoxicity by freshly isolated natural killer cells. *J. Exp. Med.* 181, 1235–1238 (1995).
15. Medvedev, A. E. *et al.* Regulation of Fas and Fas ligand expression in NK cells by cytokines and the involvement of Fas ligand in NK/LAK cell-mediated cytotoxicity. *Cytokine* 9, 394–404 (1997).
16. Moretta, A. Mechanisms in cell-mediated cytotoxicity. *Cell* 90, 13–18 (1997).
17. Tanaka, M., Itai, T., Adachi, M. & Nagata, S. Downregualtion of Fas ligand by shedding. *Nature Med.* 4, 31–36 (1998).
18. Gelmini, S. *et al.* Quantitative PCR-based homogeneous assay with fluorogenic probes to measure c-erbB-2 oncogene amplification. *Clin. Chem.* 43, 752–758 (1997).
19. Emery, J. G. *et al.* Osteoprotegerin is a receptor for the cytotoxic ligand TRAIL. *J. Biol. Chem.* 273, 14363–14367 (1998).
20. Wallach, D. Placing death under control. *Nature* 388, 123–125 (1997).
21. Colllota, F. *et al.* Interleukin-1 type II receptor: a decoy target for IL-1 that is regulated by IL-4. *Science* 261, 472–475 (1993).

22. Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* 281, 1305–1308 (1998).
23. Ashkenazi, A. & Chamow, S. M. Immunoadhesins as research tools and therapeutic agents. *Curr. Opin. Immunol.* 9, 195–200 (1997).
24. Marsters, S. *et al.* Activation of apoptosis by Apo-2 ligand is independent of FADD but blocked by CrmA. *Curr. Biol.* 6, 750–752 (1996).

Acknowledgements. We thank C. Clark, D. Pennica and V. Dixit for comments, and J. Kern and P. Quirke for tumour specimens.

Correspondence and requests for materials should be addressed to A.A. (e-mail: aa@gene.com). The GenBank accession number for the DcR3 cDNA sequence is AF104419.

Crystal structure of the ATP-binding subunit of an ABC transporter

Li-Wei Hung*, Iris Xiaoyan Wang†, Kishiko Nikaido†, Pei-Qi Liu†, Giovanna Ferro-Luzzi Ames† & Sung-Hou Kim*‡

** E. O. Lawrence Berkeley National Laboratory, † Department of Molecular and Cell Biology, and ‡ Department of Chemistry, University of California at Berkeley, Berkeley, California 94720, USA*

ABC transporters (also known as traffic ATPases) form a large family of proteins responsible for the translocation of a variety of compounds across membranes of both prokaryotes and eukaryotes[1]. The recently completed *Escherichia coli* genome sequence revealed that the largest family of paralogous *E. coli* proteins is composed of ABC transporters[2]. Many eukaryotic proteins of medical significance belong to this family, such as the cystic fibrosis transmembrane conductance regulator (CFTR), the P-glycoprotein (or multidrug-resistance protein) and the heterodimeric transporter associated with antigen processing (Tap1–Tap2). Here we report the crystal structure at 1.5 Å resolution of HisP, the ATP-binding subunit of the histidine permease, which is an ABC transporter from *Salmonella typhimurium*. We correlate the details of this structure with the biochemical, genetic and biophysical properties of the wild-type and several mutant HisP proteins. The structure provides a basis for understanding properties of ABC transporters and of defective CFTR proteins.

ABC transporters contain four structural domains: two nucleotide-binding domains (NBDs), which are highly conserved throughout the family, and two transmembrane domains[1]. In prokaryotes these domains are often separate subunits which are assembled into a membrane-bound complex; in eukaryotes the domains are generally fused into a single polypeptide chain. The periplasmic histidine permease of *S. typhimurium* and *E. coli*[1,3–8] is a well-characterized ABC transporter that is a good model for this superfamily. It consists of a membrane-bound complex, $HisQMP_2$, which comprises integral membrane subunits, HisQ and HisM, and two copies of HisP, the ATP-binding subunit. HisP, which has properties intermediate between those of integral and peripheral membrane proteins[9], is accessible from both sides of the membrane, presumably by its interaction with HisQ and HisM[6]. The two HisP subunits form a dimer, as shown by their cooperativity in ATP hydrolysis[5], the requirement for both subunits to be present for activity[8], and the formation of a HisP dimer upon chemical cross-linking. Soluble HisP also forms a dimer[3]. HisP has been purified and characterized in an active soluble form[3] which can be reconstituted into a fully active membrane-bound complex[8].

The overall shape of the crystal structure of the HisP monomer is that of an 'L' with two thick arms (arm I and arm II); the ATP-binding pocket is near the end of arm I (Fig. 1). A six-stranded β-sheet (β3 and β8–β12) spans both arms of the L, with a domain of a α- plus β-type structure (β1, β2, β4–β7, α1 and α2) on one side (within arm I) and a domain of mostly α-helices (α3–α9) on the

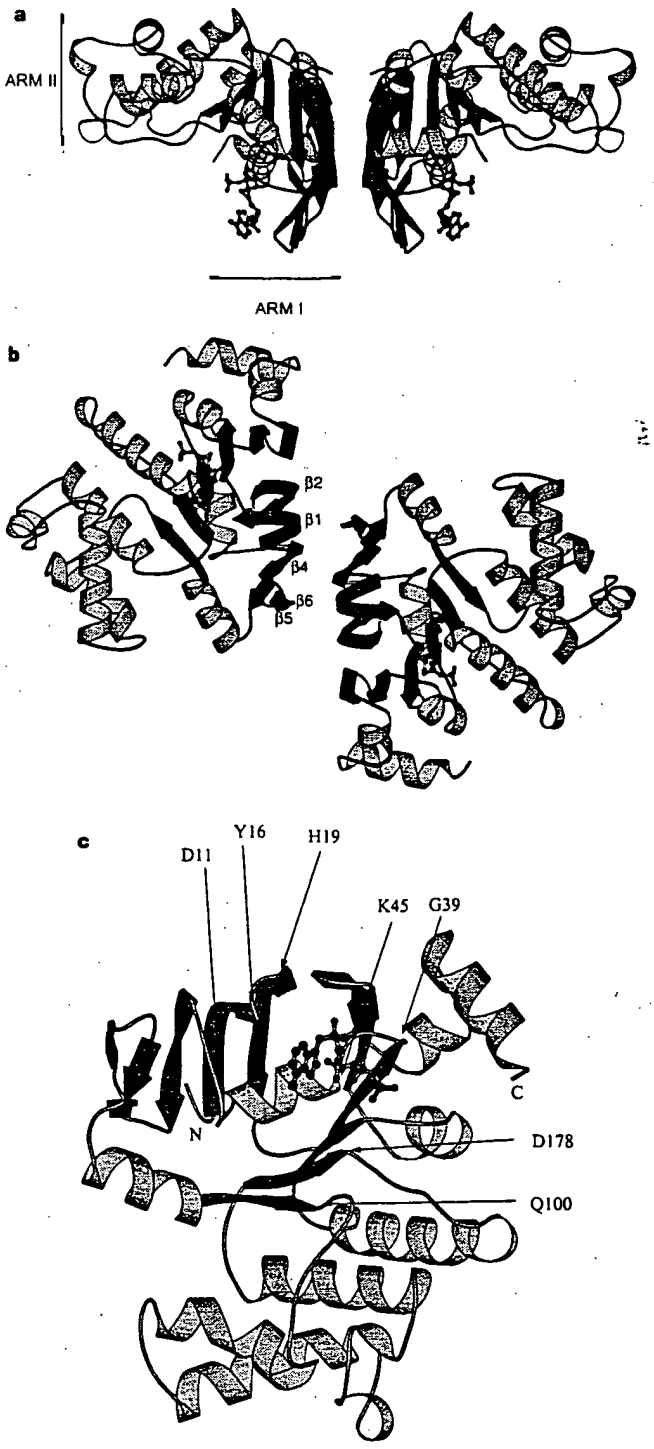


Figure 1 Crystal structure of HisP. **a**, View of the dimer along an axis perpendicular to its two-fold axis. The top and bottom of the dimer are suggested to face towards the periplasmic and cytoplasmic sides, respectively (see text). The thickness of arm II is about 25 Å, comparable to that of membrane. α-Helices are shown in orange and β-sheets in green. **b**, View along the two-fold axis of the HisP dimer, showing the relative displacement of the monomers not apparent in **a**. The β-strands at the dimer interface are labelled. **c**, View of one monomer from the bottom of arm I, as shown in **a**, towards arm II, showing the ATP-binding pocket. **a**–**c**, The protein and the bound ATP are in 'ribbon' and 'ball-and-stick' representations, respectively. Key residues discussed in the text are indicated in **c**. These figures were prepared with MOLSCRIPT[29]. N, amino terminus; C, C terminus.

Int. J. Cancer: 78, 661–666 (1998)

Publication of the International Union Against Cancer
Publication de l'Union Internationale Contre le Cancer

NOVEL APPROACH TO QUANTITATIVE POLYMERASE CHAIN REACTION USING REAL-TIME DETECTION: APPLICATION TO THE DETECTION OF GENE AMPLIFICATION IN BREAST CANCER

Ivan BIÈCHE[1,2], Martine OLIVI[1], Marie-Hélène CHAMPÈME[2], Dominique VIDAUD[1], Rosette LIDEREAU[2] and Michel VIDAUD[1]*

[1]*Laboratoire de Génétique Moléculaire, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, Paris, France*
[2]*Laboratoire d'Oncogénétique, Centre René Huguenin, St-Cloud, France*

Gene amplification is a common event in the progression of human cancers, and amplified oncogenes have been shown to have diagnostic, prognostic and therapeutic relevance. A kinetic quantitative polymerase-chain-reaction (PCR) method, based on fluorescent TaqMan methodology and a new instrument (ABI Prism 7700 Sequence Detection System) capable of measuring fluorescence in real-time, was used to quantify gene amplification in tumor DNA. Reactions are characterized by the point during cycling when PCR amplification is still in the exponential phase, rather than the amount of PCR product accumulated after a fixed number of cycles. None of the reaction components is limited during the exponential phase, meaning that values are highly reproducible in reactions starting with the same copy number. This greatly improves the precision of DNA quantification. Moreover, real-time PCR does not require post-PCR sample handling, thereby preventing potential PCR-product carry-over contamination; it possesses a wide dynamic range of quantification and results in much faster and higher sample throughput. The real-time PCR method, was used to develop and validate a simple and rapid assay for the detection and quantification of the 3 most frequently amplified genes (*myc*, *ccnd1* and *erb*B2) in breast tumors. Extra copies of *myc*, *ccnd1* and *erb*B2 were observed in 10, 23 and 15%, respectively, of 108 breast-tumor DNA; the largest observed numbers of gene copies were 4.6, 18.6 and 15.1, respectively. These results correlated well with those of Southern blotting. The use of this new semi-automated technique will make molecular analysis of human cancers simpler and more reliable, and should find broad applications in clinical and research settings. *Int. J. Cancer* 78:661–666, 1998.
© *1998 Wiley-Liss, Inc.*

Gene amplification plays an important role in the pathogenesis of various solid tumors, including breast cancer, probably because over-expression of the amplified target genes confers a selective advantage. The first technique used to detect genomic amplification was cytogenetic analysis. Amplification of several chromosome regions, visualized either as extrachromosomal double minutes (dmins) or as integrated homogeneously staining regions (HSRs), are among the main visible cytogenetic abnormalities in breast tumors. Other techniques such as comparative genomic hybridization (CGH) (Kallioniemi *et al.*, 1994) have also been used in broad searches for regions of increased DNA copy numbers in tumor cells, and have revealed some 20 amplified chromosome regions in breast tumors. Positional cloning efforts are underway to identify the critical gene(s) in each amplified region. To date, genes known to be amplified frequently in breast cancers include *myc* (8q24), *ccnd*1 (11q13), and *erb*B2 (17q12-q21) (for review, see Bièche and Lidereau, 1995).

Amplification of the *myc*, *ccnd1*, and *erb*B2 proto-oncogenes should have clinical relevance in breast cancer, since independent studies have shown that these alterations can be used to identify sub-populations with a worse prognosis (Berns *et al.*, 1992; Schuuring *et al.*, 1992; Slamon *et al.*, 1987). Muss *et al.* (1994) suggested that these gene alterations may also be useful for the prediction and assessment of the efficacy of adjuvant chemotherapy and hormone therapy.

However, published results diverge both in terms of the frequency of these alterations and their clinical value. For instance, over 500 studies in 10 years have failed to resolve the controversy surrounding the link suggested by Slamon *et al.* (1987) between *erb*B2 amplification and disease progression. These discrepancies are partly due to the clinical, histological and ethnic heterogeneity of breast cancer, but technical considerations are also probably involved.

Specific genes (DNA) were initially quantified in tumor cells by means of blotting procedures such as Southern and slot blotting. These batch techniques require large amounts of DNA (5–10 µg/reaction) to yield reliable quantitative results. Furthermore, meticulous care is required at all stages of the procedures to generate blots of sufficient quality for reliable dosage analysis. Recently, PCR has proven to be a powerful tool for quantitative DNA analysis, especially with minimal starting quantities of tumor samples (small, early-stage tumors and formalin-fixed, paraffin-embedded tissues).

Quantitative PCR can be performed by evaluating the amount of product either after a given number of cycles (end-point quantitative PCR) or after a varying number of cycles during the exponential phase (kinetic quantitative PCR). In the first case, an internal standard distinct from the target molecule is required to ascertain PCR efficiency. The method is relatively easy but implies generating, quantifying and storing an internal standard for each gene studied. Nevertheless, it is the most frequently applied method to date.

One of the major advantages of the kinetic method is its rapidity in quantifying a new gene, since no internal standard is required (an external standard curve is sufficient). Moreover, the kinetic method has a wide dynamic range (at least 5 orders of magnitude), giving an accurate value for samples differing in their copy number. Unfortunately, the method is cumbersome and has therefore been rarely used. It involves aliquot sampling of each assay mix at regular intervals and quantifying, for each aliquot, the amplification product. Interest in the kinetic method has been stimulated by a novel approach using fluorescent TaqMan methodology and a new instrument (ABI Prism 7700 Sequence Detection System) capable of measuring fluorescence in real time (Gibson *et al.*, 1996; Heid *et al.*, 1996). The TaqMan reaction is based on the 5′ nuclease assay first described by Holland *et al.* (1991). The latter uses the 5′ nuclease activity of Taq polymerase to cleave a specific fluorogenic oligonucleotide probe during the extension phase of PCR. The approach uses dual-labeled fluorogenic hybridization probes (Lee *et al.*, 1993). One fluorescent dye, co-valently linked to the 5′ end of the oligonucleotide, serves as a reporter [FAM (*i.e.*, 6-carboxy-fluorescein)] and its emission spectrum is quenched by a second fluorescent dye, TAMRA (*i.e.*, 6-carboxy-tetramethyl-rhodamine) attached to the 3′ end. During the extension phase of the PCR

Grant sponsors: Association Pour la Recherche sur le Cancer and Ministère de l'Enseignement Supérieur et de la Recherche.

*Correspondence to: Laboratoire de Génétique Moléculaire, Faculté des Sciences Pharmaceutiques et Biologiques de Paris, 4 Avenue de l'Observatoire, F-75006 Paris, France. Fax: (33)1-4407-1754.
E-mail: mvidaud@teaser.fr

Received 2 May 1998; Revised 30 June 1998

cycle, the fluorescent hybridization probe is hydrolyzed by the 5'-3' nucleolytic activity of DNA polymerase. Nuclease degradation of the probe releases the quenching of FAM fluorescence emission, resulting in an increase in peak fluorescence emission. The fluorescence signal is normalized by dividing the emission intensity of the reporter dye (FAM) by the emission intensity of a reference dye (*i.e.*, ROX, 6-carboxy-X-rhodamine) included in TaqMan buffer, to obtain a ratio defined as the Rn (normalized reporter) for a given reaction tube. The use of a sequence detector enables the fluorescence spectra of all 96 wells of the thermal cycler to be measured continuously during PCR amplification.

The real-time PCR method offers several advantages over other current quantitative PCR methods (Celi *et al.*, 1994): *(i)* the probe-based homogeneous assay provides a real-time method for detecting only specific amplification products, since specific hybridation of both the primers and the probe is necessary to generate a signal; *(ii)* the C_t (threshold cycle) value used for quantification is measured when PCR amplification is still in the log phase of PCR product accumulation. This is the main reason why C_t is a more reliable measure of the starting copy number than are end-point measurements, in which a slight difference in a limiting component can have a drastic effect on the amount of product; *(iii)* use of C_t values gives a wider dynamic range (at least 5 orders of magnitude), reducing the need for serial dilution; *(iv)* The real-time PCR method is run in a closed-tube system and requires no post-PCR sample handling, thus avoiding potential contamination; *(v)* the system is highly automated, since the instrument continuously measures fluorescence in all 96 wells of the thermal cycler during PCR amplification and the corresponding software processes, and analyzes the fluorescence data; *(vi)* the assay is rapid, as results are available just one minute after thermal cycling is complete; *(vii)* the sample throughput of the method is high, since 96 reactions can be analyzed in 2 hr.

Here, we applied this semi-automated procedure to determine the copy numbers of the 3 most frequently amplified genes in breast tumors (*myc*, *ccnd1* and *erb*B2), as well as 2 genes (*alb* and *app*) located in a chromosome region in which no genetic changes have been observed in breast tumors. The results for 108 breast tumors were compared with previous Southern-blot data for the same samples.

MATERIAL AND METHODS

Tumor and blood samples

Samples were obtained from 108 primary breast tumors removed surgically from patients at the Centre René Huguenin; none of the patients had undergone radiotherapy or chemotherapy. Immediately after surgery, the tumor samples were placed in liquid nitrogen until extraction of high-molecular-weight DNA. Patients were included in this study if the tumor sample used for DNA preparation contained more than 60% of tumor cells (histological analysis). A blood sample was also taken from 18 of the same patients.

DNA was extracted from tumor tissue and blood leukocytes according to standard methods.

Real-time PCR

Theoretical basis. Reactions are characterized by the point during cycling when amplification of the PCR product is first detected, rather than by the amount of PCR product accumulated after a fixed number of cycles. The higher the starting copy number of the genomic DNA target, the earlier a significant increase in fluorescence is observed. The parameter C_t (threshold cycle) is defined as the fractional cycle number at which the fluorescence generated by cleavage of the probe passes a fixed threshold above baseline. The target gene copy number in unknown samples is quantified by measuring C_t and by using a standard curve to determine the starting copy number. The precise amount of genomic DNA (based on optical density) and its quality (*i.e.*, lack of extensive degradation) are both difficult to assess. We therefore also quantified a control gene *(alb)* mapping to chromosome region 4q11-q13, in which no genetic alterations have been found in breast-tumor DNA by means of CGH (Kallioniemi *et al.*, 1994).

Thus, the ratio of the copy number of the target gene to the copy number of the *alb* gene normalizes the amount and quality of genomic DNA. The ratio defining the level of amplification is termed "N", and is determined as follows:

$$N = \frac{\text{copy number of target gene } (app,\ myc,\ ccnd1,\ erbB2)}{\text{copy number of reference gene } (alb)}.$$

Primers, probes, reference human genomic DNA and PCR consumables. Primers and probes were chosen with the assistance of the computer programs Oligo 4.0 (National Biosciences, Plymouth, MN), EuGene (Daniben Systems, Cincinnati, OH) and Primer Express (Perkin-Elmer Applied Biosystems, Foster City, CA).

Primers were purchased from DNAgency (Malvern, PA) and probes from Perkin-Elmer Applied Biosystems.

Nucleotide sequences for the oligonucleotide hybridization probes and primers are available on request.

The TaqMan PCR Core reagent kit, MicroAmp optical tubes, and MicroAmp caps were from Perkin-Elmer Applied Biosystems.

Standard-curve construction. The kinetic method requires a standard curve. The latter was constructed with serial dilutions of specific PCR products, according to Piatak *et al.* (1993). In practice, each specific PCR product was obtained by amplifying 20 ng of a standard human genomic DNA (Boehringer, Mannheim, Germany) with the same primer pairs as those used later for real-time quantitative PCR. The 5 PCR products were purified using MicroSpin S-400 HR columns (Pharmacia, Uppsala, Sweden) electrophorezed through an acrylamide gel and stained with ethidium bromide to check their quality. The PCR products were then quantified spectrophotometrically and pooled, and serially diluted 10-fold in mouse genomic DNA (Clontech, Palo Alto, CA) at a constant concentration of 2 ng/µl. The standard curve used for real-time quantitative PCR was based on serial dilutions of the pool of PCR products ranging from 10^{-7} (10^5 copies of each gene) to 10^{-10} (10^2 copies). This series of diluted PCR products was aliquoted and stored at −80°C until use.

The standard curve was validated by analyzing 2 known quantities of calibrator human genomic DNA (20 ng and 50 ng).

PCR amplification. Amplification mixes (50 µl) contained the sample DNA (around 20 ng, around 6600 copies of disomic genes), 10× TaqMan buffer (5 µl), 200 µM dATP, dCTP, dGTP, and 400 µM dUTP, 5 mM $MgCl_2$, 1.25 units of AmpliTaq Gold, 0.5 units of AmpErase uracil N-glycosylase (UNG), 200 nM each primer and 100 nM probe. The thermal cycling conditions comprised 2 min at 50°C and 10 min at 95°C. Thermal cycling consisted of 40 cycles at 95°C for 15 s and 65°C for 1 min. Each assay included: a standard curve (from 10^5 to 10^2 copies) in duplicate, a no-template control, 20 ng and 50 ng of calibrator human genomic DNA (Boehringer) in triplicate, and about 20 ng of unknown genomic DNA in triplicate (26 samples can thus be analyzed on a 96-well microplate). All samples with a coefficient of variation (CV) higher than 10% were retested.

All reactions were performed in the ABI Prism 7700 Sequence Detection System (Perkin-Elmer Applied Biosystems), which detects the signal from the fluorogenic probe during PCR.

Equipment for real-time detection. The 7700 system has a built-in thermal cycler and a laser directed via fiber optical cables to each of the 96 sample wells. A charge-coupled-device (CDD) camera collects the emission from each sample and the data are analyzed automatically. The software accompanying the 7700 system calculates C_t and determines the starting copy number in the samples.

Determination of gene amplification. Gene amplification was calculated as described above. Only samples with an N value higher than 2 were considered to be amplified.

RESULTS

To validate the method, real-time PCR was performed on genomic DNA extracted from 108 primary breast tumors, and 18 normal leukocyte DNA samples from some of the same patients. The target genes were the *myc*, *ccnd1* and *erb*B2 proto-oncogenes, and the β-amyloid precursor protein gene (*app*), which maps to a chromosome region (21q21.2) in which no genetic alterations have been found in breast tumors (Kallioniemi *et al.*, 1994). The reference disomic gene was the albumin gene (*alb*, chromosome 4q11-q13).

Validation of the standard curve and dynamic range of real-time PCR

The standard curve was constructed from PCR products serially diluted in genomic mouse DNA at a constant concentration of 2 ng/μl. It should be noted that the 5 primer pairs chosen to analyze the 5 target genes do not amplify genomic mouse DNA (data not shown). Figure 1 shows the real-time PCR standard curve for the *alb* gene. The dynamic range was wide (at least 4 orders of magnitude), with samples containing as few as 10^2 copies or as many as 10^5 copies.

*Copy-number ratio of the 2 reference genes (*app *and* alb*)*

The *app* to *alb* copy-number ratio was determined in 18 normal leukocyte DNA samples and all 108 primary breast-tumor DNA

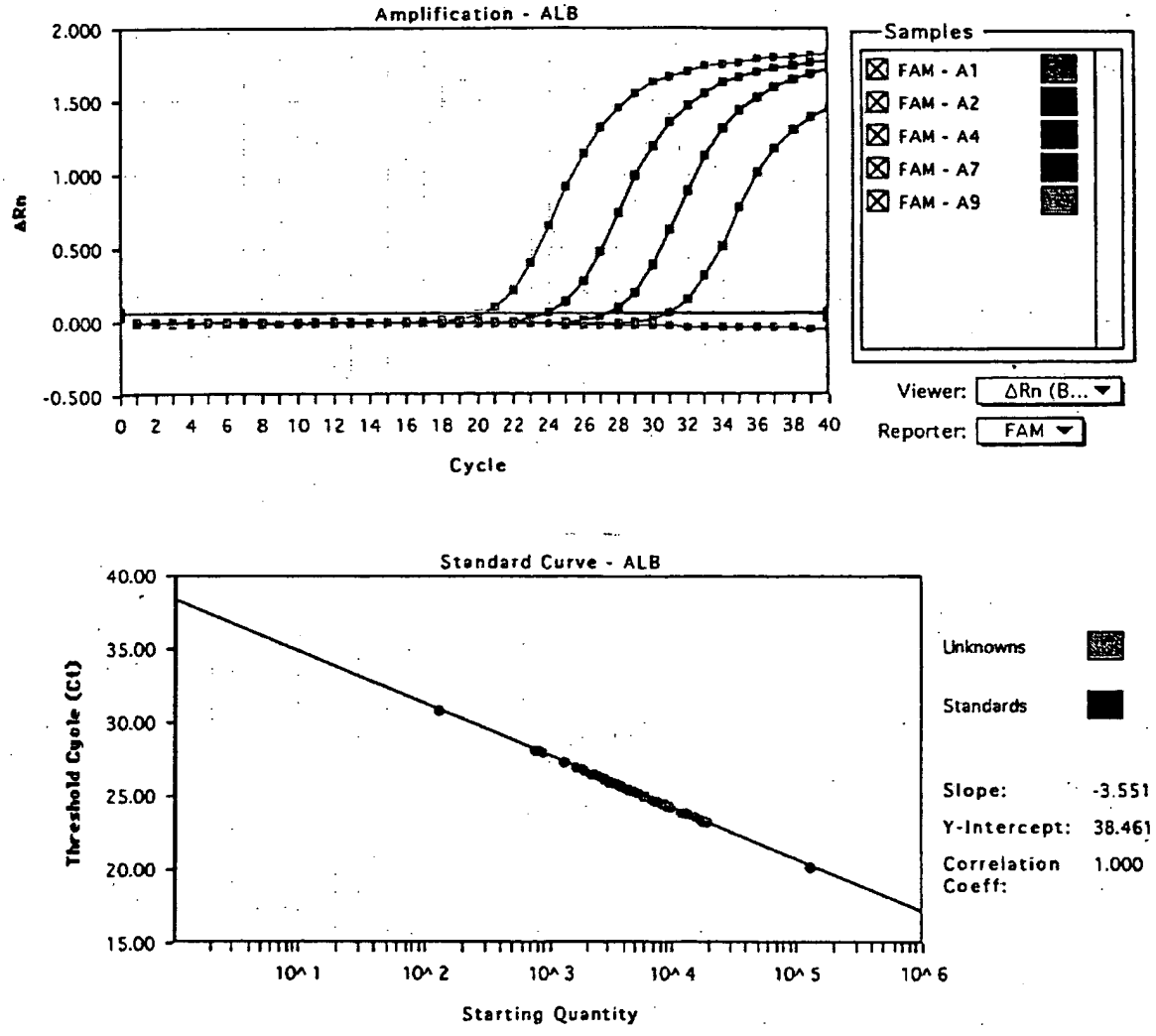


FIGURE 1 – Albumin (*alb*) gene dosage by real-time PCR. Top: Amplification plots for reactions with starting *alb* gene copy number ranging from 10^5 (A9), 10^4 (A7), 10^3 (A4) to 10^2 (A2) and a no-template control (A1). Cycle number is plotted *vs.* change in normalized reporter signal (ΔRn). For each reaction tube, the fluorescence signal of the reporter dye (FAM) is divided by the fluorescence signal of the passive reference dye (ROX), to obtain a ratio defined as the normalized reporter signal (Rn). ΔRn represents the normalized reporter signal (Rn) minus the baseline signal established in the first 15 PCR cycles. ΔRn increases during PCR as *alb* PCR product copy number increases until the reaction reaches a plateau. C_t (threshold cycle) represents the fractional cycle number at which a significant increase in Rn above a baseline signal (horizontal black line) can first be detected. Two replicate plots were performed for each standard sample, but the data for only one are shown here. Bottom: Standard curve plotting log starting copy number *vs.* C_t (threshold cycle). The black dots represent the data for standard samples plotted in duplicate and the red dots the data for unknown genomic DNA samples plotted in triplicate. The standard curve shows 4 orders of linear dynamic range.

samples. We selected these 2 genes because they are located in 2 chromosome regions (*app*, 21q21.2; *alb*, 4q11-q13) in which no obvious genetic changes (including gains or losses) have been observed in breast cancers (Kallioniemi *et al.*, 1994). The ratio for the 18 normal leukocyte DNA samples fell between 0.7 and 1.3 (mean 1.02 ± 0.21), and was similar for the 108 primary breast-tumor DNA samples (0.6 to 1.6, mean 1.06 ± 0.25), confirming that *alb* and *app* are appropriate reference disomic genes for breast-tumor DNA. The low range of the ratios also confirmed that the nucleotide sequences chosen for the primers and probes were not polymorphic, as mismatches of their primers or probes with the subject's DNA would have resulted in differential amplification.

myc, ccnd1 *and* erb*B2 gene dose in normal leukocyte DNA*

To determine the cut-off point for gene amplification in breast-cancer tissue, 18 normal leukocyte DNA samples were tested for the gene dose (N), calculated as described in "Material and Methods". The N value of these samples ranged from 0.5 to 1.3 (mean 0.84 ± 0.22) for *myc*; 0.7 to 1.6 (mean 1.06 ± 0.23) for *ccnd1* and 0.6 to 1.3 (mean 0.91 ± 0.19) for *erb*B2. Since N values for *myc*, *ccnd1* and *erb*B2 in normal leukocyte DNA consistently fell between 0.5 and 1.6, values of 2 or more were considered to represent gene amplification in tumor DNA.

myc, ccnd1 *and* erb*B2 gene dose in breast-tumor DNA*

myc, *ccnd1* and *erb*B2 gene copy numbers in the 108 primary breast tumors are reported in Table I. Extra copies of *ccnd1* were more frequent (23%, 25/108) than extra copies of *erb*B2 (15%, 16/108) and *myc* (10%, 11/108), and ranged from 2 to 18.6 for *ccnd1*, 2 to 15.1 for *erb*B2, and only 2 to 4.6 for the *myc* gene. Figure 2 and Table II represent tumors in which the *ccnd1* gene was amplified 16-fold (T145), 6-fold (T133) and non-amplified (T118). The 3 genes were never found to be co-amplified in the same tumor. *erb*B2 and *ccnd1* were co-amplified in only 3 cases, *myc* and *ccnd1* in 2 cases and *myc* and *erb*B2 in 1 case. This favors the hypothesis that gene amplifications are independent events in breast cancer. Interestingly, 5 tumors showed a decrease of at least 50% in the *erb*B2 copy number ($N < 0.5$), suggesting that they bore deletions of the 17q21 region (the site of *erb*B2). No such decrease in copy number was observed with the other 2 proto-oncogenes.

Comparison of gene dose determined by real-time quantitative PCR and Southern-blot analysis

Southern-blot analysis of *myc*, *ccnd1* and *erb*B2 amplifications had previously been done on the same 108 primary breast tumors. A perfect correlation between the results of real-time PCR and Southern blot was obtained for tumors with high copy numbers ($N \geq 5$). However, there were cases (1 *myc*, 6 *ccnd1* and 4 *erb*B2) in which real-time PCR showed gene amplification whereas Southern-blot did not, but these were mainly cases with low extra copy numbers (N from 2 to 2.9).

DISCUSSION

The clinical applications of gene amplification assays are currently limited, but would certainly increase if a simple, standardized and rapid method were perfected. Gene amplification status has been studied mainly by means of Southern blotting, but this method is not sensitive enough to detect low-level gene amplification nor accurate enough to quantify the full range of amplification values. Southern blotting is also time-consuming, uses radioactive reagents and requires relatively large amounts of high-quality genomic DNA, which means it cannot be used routinely in many laboratories. An amplification step is therefore required to determine the copy number of a given target gene from minimal quantities of tumor DNA (small early-stage tumors, cytopuncture specimens or formalin-fixed, paraffin-embedded tissues).

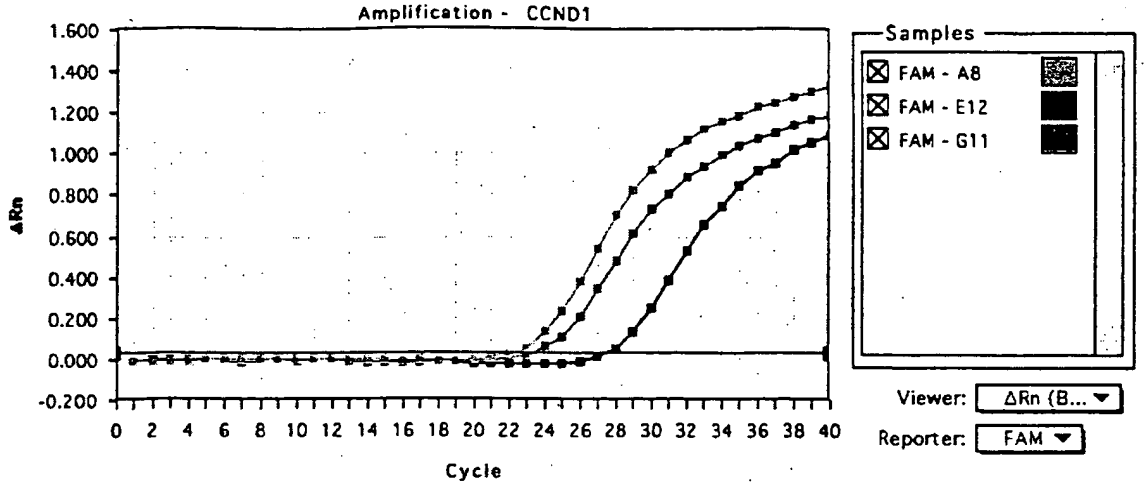
TABLE I – DISTRIBUTION OF AMPLIFICATION LEVEL (N) FOR *myc*, *ccnd1* AND *erb*B2 GENES IN 108 HUMAN BREAST TUMORS

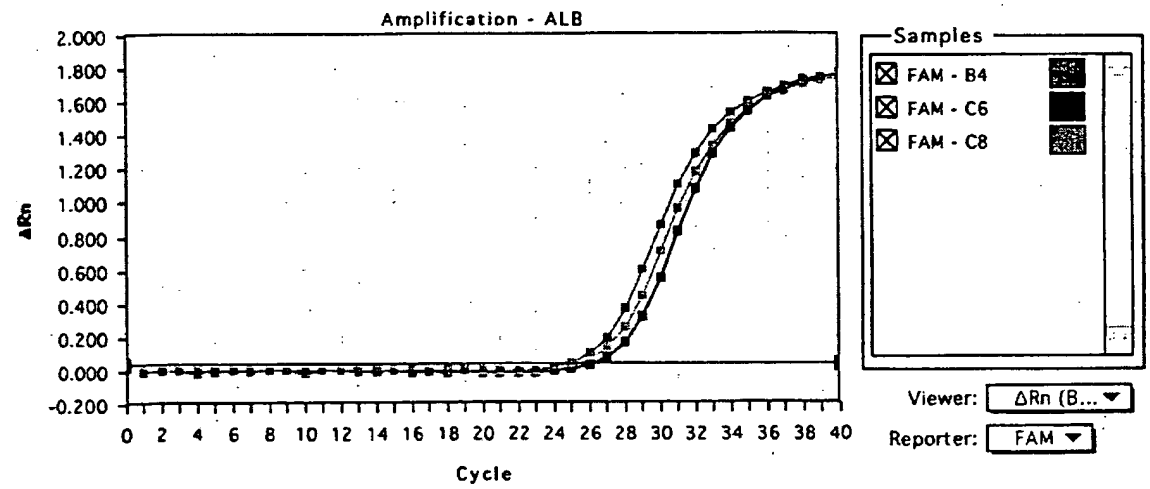
Gene	Amplification level (N)			
	<0.5	0.5–1.9	2–4.9	≥5
myc	0	97 (89.8%)	11 (10.2%)	0
ccnd1	0	83 (76.9%)	17 (15.7%)	8 (7.4%)
*erb*B2	5 (4.6%)	87 (80.6%)	8 (7.4%)	8 (7.4%)

In this study, we validated a PCR method developed for the quantification of gene over-representation in tumors. The method, based on real-time analysis of PCR amplification, has several advantages over other PCR-based quantitative assays such as competitive quantitative PCR (Celi *et al.*, 1994). First, the real-time PCR method is performed in a closed-tube system, avoiding the risk of contamination by amplified products. Re-amplification of carryover PCR products in subsequent experiments can also be prevented by using the enzyme uracil N-glycosylase (UNG) (Longo *et al.*, 1990). The second advantage is the simplicity and rapidity of sample analysis, since no post-PCR manipulations are required. Our results show that the automated method is reliable. We found it possible to determine, in triplicate, the number of copies of a target gene in more than 100 tumors per day. Third, the system has a linear dynamic range of at least 4 orders of magnitude, meaning that samples do not have to contain equal starting amounts of DNA. This technique should therefore be suitable for analyzing formalin-fixed, paraffin-embedded tissues. Fourth, and above all, real-time PCR makes DNA quantification much more precise and reproducible, since it is based on C_t values rather than end-point measurement of the amount of accumulated PCR product. Indeed, the ABI Prism 7700 Sequence Detection System enables C_t to be calculated when PCR amplification is still in the exponential phase and when none of the reaction components is rate-limiting. The within-run CV of the C_t value for calibrator human DNA (5 replicates) was always below 5%, and the between-assay precision in 5 different runs was always below 10% (data not shown). In addition, the use of a standard curve is not absolutely necessary, since the copy number can be determined simply by comparing the C_t ratio of the target gene with that of reference genes. The results obtained by the 2 methods (with and without a standard curve) are similar in our experiments (data not shown). Moreover, unlike competitive quantitative PCR, real-time PCR does not require an internal control (the design and storage of internal controls and the validation of their amplification efficiency is laborious).

The only potential disavantage of real-time PCR, like all other PCR-based methods and solid-matrix blotting techniques (Southern blots and dot blots) is that is cannot avoid dilution artifacts inherent in the extraction of DNA from tumor cells contained in heterogeneous tissue specimens. Only FISH and immunohistochemistry can measure alterations on a cell-by-cell basis (Pauletti *et al.*, 1996; Slamon *et al.*, 1989). However, FISH requires expensive equipment and trained personnel and is also time-consuming. Moreover, FISH does not assess gene expression and therefore cannot detect cases in which the gene product is over-expressed in the absence of gene amplification, which will be possible in the future by real-time quantitative RT-PCR. Immunohistochemistry is subject to considerable variations in the hands of different teams, owing to alterations of target proteins during the procedure, the different primary antibodies and fixation methods used and the criteria used to define positive staining.

The results of this study are in agreement with those reported in the literature. (*i*) Chromosome regions 4q11-q13 and 21q21.2 (which bear *alb* and *app*, respectively) showed no genetic alterations in the breast-cancer samples studied here, in keeping with the results of CGH (Kallioniemi *et al.*, 1994). (*ii*) We found that amplifications of these 3 oncogenes were independent events, as reported by other teams (Berns *et al.*, 1992; Borg *et al.*, 1992). (*iii*) The frequency and degree of *myc* amplification in our breast tumor DNA series were lower than those of *ccnd1* and *erb*B2 amplification, confirming the findings of Borg *et al.* (1992) and Courjal *et al.* (1997). (*iv*) The maxima of *ccnd1* and *erb*B2 over-representation were 18-fold and 15-fold, also in keeping with earlier results (about





Tumor	CCND1 C_t	CCND1 Copy number	ALB C_t	ALB Copy number
T118	27.3	4605	26.5	4365
T133	23.2	61659	25.2	10092
T145	22.1	125892	25.6	7762

FIGURE 2 – *ccnd1* and alb gene dosage by real-time PCR in 3 breast tumor samples: T118 (E12, C6, black squares), T133 (G11, B4, red squares) and T145 (A8, C8, blue squares). Given the C_t of each sample, the initial copy number is inferred from the standard curve obtained during the same experiment. Triplicate plots were performed for each tumor sample, but the data for only one are shown here. The results are shown in Table II.

30-fold maximum) (Berns *et al.*, 1992; Borg *et al.*, 1992; Courjal *et al.*, 1997). (*v*) The *erb*B2 copy numbers obtained with real-time PCR were in good agreement with data obtained with other quantitative PCR-based assays in terms of the frequency and degree of amplification (An *et al.*, 1995; Deng *et al.*, 1996; Valeron *et al.*, 1996). Our results also correlate well with those recently published by Gelmini *et al.* (1997), who used the TaqMan system to measure *erb*B2 amplification in a small series of breast tumors (n = 25), but with an instrument (LS-50B luminescence spectrometer, Perkin-Elmer Applied Biosystems) which only allows end-

TABLE II - EXAMPLES OF *ccnd1* GENE DOSAGE RESULTS FROM 3 BREAST TUMORS[1]

Tumor	*ccnd1*			*alb*			Nccnd1/alb
	Copy number	Mean	SD	Copy number	Mean	SD	
T118	4525			4223			
	4605	4603	77	4365	4325	89	1.06
	4678			4387			
T133	59821			9787			
	61659	61100	1111	10092	10137	375	6.03
	61821			10533			
T145	128563			7321			
	125892	125392	3448	7762	7672	316	16.34
	121722			7933			

[1]For each sample, 3 replicate experiments were performed and the mean and the standard deviation (SD) was determined. The level of *ccnd1* gene amplification (N*ccnd1/alb*) is determined by dividing the average *ccnd1* copy number value by the average *alb* copy number value.

point measurement of fluorescence intensity. Here we report *myc* and *ccnd1* gene dosage in breast cancer by means of quantitative PCR. (vi) We found a high degree of concordance between real-time quantitative PCR and Southern blot analysis in terms of gene amplification, especially for samples with high copy numbers (≥5-fold). The slightly higher frequency of gene amplification (especially *ccnd1* and *erb*B2) observed by means of real-time quantitative PCR as compared with Southern-blot analysis may be explained by the higher sensitivity of the former method. However, we cannot rule out the possibility that some tumors with a few extra gene copies observed in real-time PCR had additional copies of an arm or a whole chromosome (trisomy, tetrasomy or polysomy) rather than true gene amplification. These 2 types of genetic alteration (polysomy and gene amplification) could be easily distinguished in the future by using an additional probe located on the same chromosome arm, but some distance from the target gene. It is noteworthy that high gene copy numbers have the greatest prognostic significance in breast carcinoma (Borg *et al.*, 1992; Slamon *et al.*, 1987).

Finally, this technique can be applied to the detection of gene deletion as well as gene amplification. Indeed, we found a decreased copy number of *erb*B2 (but not of the other 2 proto-oncogenes) in several tumors; *erb*B2 is located in a chromosome region (17q21) reported to contain both deletions and amplifications in breast cancer (Bièche and Lidereau, 1995).

In conclusion, gene amplification in various cancers can be used as a marker of pre-neoplasia, also for early diagnosis of cancer, staging, prognostication and choice of treatment. Southern blotting is not sufficiently sensitive, and FISH is lengthy and complex. Real-time quantitative PCR overcomes both these limitations, and is a sensitive and accurate method of analyzing large numbers of samples in a short time. It should find a place in routine clinical gene dosage.

ACKNOWLEDGEMENTS

RL is a research director at the Institut National de la Santé et de la Recherche Médicale (INSERM). We thank the staff of the Centre René Huguenin for assistance in specimen collection and patient care.

REFERENCES

AN, H.X., NIEDERACHER, D., BECKMANN, M.W., GÖHRING, U.J., SCHARL, A., PICARD, F., VAN ROEYEN, C., SCHNÜRCH, H.G. and BENDER, H.G., *erb*B2 gene amplification detected by fluorescent differential polymerase chain reaction in paraffin-embedded breast carcinoma tissues. *Int. J. Cancer (Pred. Oncol.)*, **64**, 291–297 (1995).

BERNS, E.M.J.J., KLIJN, J.G.M., VAN PUTTEN, W.L.J., VAN STAVEREN, I.L., PORTENGEN, H. and FOEKENS, J.A., c-*myc* amplification is a better prognostic factor than HER2/*neu* amplification in primary breast cancer. *Cancer Res.*, **52**, 1107–1113 (1992).

BIÈCHE, I. and LIDEREAU, R., Genetic alterations in breast cancer. *Genes Chrom. Cancer*, **14**, 227–251 (1995).

BORG, A., BALDETORP, B., FERNO, M., OLSSON, H. and SIGURDSSON, H., c-*myc* amplification is an independent prognostic factor in post-menopausal breast cancer. *Int. J. Cancer*, **51**, 687–691 (1992).

CELI, F.S., COHEN, M.M., ANTONARAKIS, S.E., WERTHEIMER, E., ROTH, J. and SHULDINER, A.R., Determination of gene dosage by a quantitative adaptation of the polymerase chain reaction (gd-PCR): rapid detection of deletions and duplications of gene sequences. *Genomics*, **21**, 304–310 (1994).

COURJAL, F., CUNY, M., SIMONY-LAFONTAINE, J., LOUASSON, G., SPEISER, P., ZEILLINGER, R., RODRIGUEZ, C. and THEILLET, C., Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups. *Cancer Res.*, **57**, 4360–4367 (1997).

DENG, G., YU, M., CHEN, L.C., MOORE, D., KURISU, W., KALLIONIEMI, A., WALDMAN, F.M., COLLINS, C. and SMITH, H.S., Amplifications of oncogene *erb*B-2 and chromosome 20q in breast cancer determined by differentially competitive polymerase chain reaction. *Breast Cancer Res. Treat.*, **40**, 271–281 (1996).

GELMINI, S., ORIANDO, C., SESTINI, R., VONA, G., PINZANI, P., RUOCCO, L. and PAZZAGLI, M., Quantitative polymerase chain reaction-based homogeneous assay with fluorogenic probes to measure c-*erB*-2 oncogene amplification. *Clin. Chem.*, **43**, 752–758 (1997).

GIBSON, U.E.M., HEID, C.A. and WILLIAMS, P.M., A novel method for real-time quantitative RT-PCR. *Genome Res.*, **6**, 995–1001 (1996).

HEID, C.A., STEVENS, J., LIVAK, K.J. and WILLIAMS, P.M., Real-time quantitative PCR. *Genome Res.*, **6**, 986–994 (1996).

HOLLAND, P.M., ABRAMSON, R.D., WATSON, R. and GELFAND, D.H., Detection of specific polymerase chain reaction product by utilizing the 5' to 3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. nat. Acad. Sci. (Wash.)*, **88**, 7276–7280 (1991).

KALLIONIEMI, A., KALLIONIEMI, O.P., PIPER, J., TANNER, M., STOKKES, T., CHEN, L., SMITH, H.S., PINKEL, D., GRAY, J.W. and WALDMAN, F.M., Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. nat. Acad. Sci. (Wash.)*, **91**, 2156–2160 (1994).

LEE, L.G., CONNELL, C.R. and BIOCH, W., Allelic discrimination by nick-translation PCR with fluorogenic probe. *Nucleic Acids Res.*, **21**, 3761–3766 (1993).

LONGO, N., BERNINGER, N.S. and HARTLEY, J.L., Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions. *Gene*, **93**, 125–128 (1990).

MUSS, H.B., THOR, A.D., BERRY, D.A., KUTE, T., LIU, E.T., KOERNER, F., CIRRINCIONE, C.T., BUDMAN, D.R., WOOD, W.C., BARCOS, M. and HENDERSON, I.C., c-*erb*B-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *New Engl. J. Med.*, **330**, 1260–1266 (1994).

PAULETTI, G., GODOLPHIN, W., PRESS, M.F. and SALMON, D.J., Detection and quantification of HER-2/*neu* gene amplification in human breast cancer archival material using fluorescence *in situ* hybridization. *Oncogene*, **13**, 63–72 (1996).

PIATAK, M., LUK, K.C., WILLIAMS, B. and LIFSON, J.D., Quantitative competitive polymerase chain reaction for accurate quantitation of HIV DNA and RNA species. *Biotechniques*, **14**, 70–80 (1993).

SCHUURING, E., VERHOEVEN, E., VAN TINTEREN, H., PETERSE, J.L., NUNNIK, B., THUNNISSEN, F.B.J.M., DEVILEE, P., CORNELISSE, C.J., VAN DE VIJVER, M.J., MOOI, W.J. and MICHALIDES, R.J.A.M., Amplification of genes within the chromosome 11q13 region is indicative of poor prognosis in patients with operable breast cancer. *Cancer Res.*, **52**, 5229–5234 (1992).

SLAMON, D.J., CLARK, G.M., WONG, S.G., LEVIN, W.S., ULLRICH, A. and MCGUIRE, W.L., Human breast cancer: correlation of relapse and survival with amplification of the HER-2/*neu* oncogene. *Science*, **235**, 177–182 (1987).

SLAMON, D.J., GODOLPHIN, W., JONES, L.A., HOLT, J.A., WONG, S.G., KEITH, D.E., LEVIN, W.J., STUART, S.G., UDOVE, J., ULLRICH, A. and PRESS, M.F., Studies of the HER-2/*neu* proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707–712 (1989).

VALERON, P.F., CHIRINO, R., FERNANDEZ, L., TORRES, S., NAVARRO, D., AGUIAR, J., CABRERA, J.J., DIAZ-CHICO, B.N. and DIAZ-CHICO, J.C., Validation of a differential PCR and an ELISA procedure in studying HER-2/*neu* status in breast cancer. *Int. J. Cancer*, **65**, 129–133 (1996).

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14717–14722, December 1998
Cell Biology, Medical Sciences

WISP genes are members of the connective tissue growth factor family that are up-regulated in Wnt-1-transformed cells and aberrantly expressed in human colon tumors

DIANE PENNICA*†, TODD A. SWANSON*, JAMES W. WELSH*, MARGARET A. ROY‡, DAVID A. LAWRENCE*, JAMES LEE‡, JENNIFER BRUSH‡, LISA A. TANEYHILL§, BETHANNE DEUEL‡, MICHAEL LEW¶, COLIN WATANABE‖, ROBERT L. COHEN*, MONA F. MELHEM**, GENE G. FINLEY**, PHIL QUIRKE††, AUDREY D. GODDARD‡, KENNETH J. HILLAN¶, AUSTIN L. GURNEY‡, DAVID BOTSTEIN‡,‡‡, AND ARNOLD J. LEVINE§

Departments of *Molecular Oncology, ‡Molecular Biology, ‖Scientific Computing, and ¶Pathology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080; **University of Pittsburgh School of Medicine, Veterans Administration Medical Center, Pittsburgh, PA 15240; ††University of Leeds, Leeds, LS29JT United Kingdom; ‡‡Department of Genetics, Stanford University, Palo Alto, CA 94305; and §Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Contributed by David Botstein and Arnold J. Levine, October 21, 1998

ABSTRACT **Wnt family members are critical to many developmental processes, and components of the Wnt signaling pathway have been linked to tumorigenesis in familial and sporadic colon carcinomas. Here we report the identification of two genes, *WISP-1* and *WISP-2*, that are up-regulated in the mouse mammary epithelial cell line C57MG transformed by Wnt-1, but not by Wnt-4. Together with a third related gene, *WISP-3*, these proteins define a subfamily of the connective tissue growth factor family. Two distinct systems demonstrated *WISP* induction to be associated with the expression of Wnt-1. These included (*i*) C57MG cells infected with a Wnt-1 retroviral vector or expressing Wnt-1 under the control of a tetracyline repressible promoter, and (*ii*) Wnt-1 transgenic mice. The *WISP-1* gene was localized to human chromosome 8q24.1–8q24.3. *WISP-1* genomic DNA was amplified in colon cancer cell lines and in human colon tumors and its RNA overexpressed (2- to >30-fold) in 84% of the tumors examined compared with patient-matched normal mucosa. *WISP-3* mapped to chromosome 6q22–6q23 and also was overexpressed (4- to >40-fold) in 63% of the colon tumors analyzed. In contrast, *WISP-2* mapped to human chromosome 20q12–20q13 and its DNA was amplified, but RNA expression was reduced (2- to >30-fold) in 79% of the tumors. These results suggest that the *WISP* genes may be downstream of Wnt-1 signaling and that aberrant levels of *WISP* expression in colon cancer may play a role in colon tumorigenesis.**

Wnt-1 is a member of an expanding family of cysteine-rich, glycosylated signaling proteins that mediate diverse developmental processes such as the control of cell proliferation, adhesion, cell polarity, and the establishment of cell fates (1, 2). Wnt-1 originally was identified as an oncogene activated by the insertion of mouse mammary tumor virus in virus-induced mammary adenocarcinomas (3, 4). Although Wnt-1 is not expressed in the normal mammary gland, expression of Wnt-1 in transgenic mice causes mammary tumors (5).

In mammalian cells, Wnt family members initiate signaling by binding to the seven-transmembrane spanning Frizzled receptors and recruiting the cytoplasmic protein Dishevelled (Dsh) to the cell membrane (1, 2, 6). Dsh then inhibits the kinase activity of the normally constitutively active glycogen synthase kinase-3β (GSK-3β) resulting in an increase in β-catenin levels. Stabilized β-catenin interacts with the transcription factor TCF/Lef1, forming a complex that appears in the nucleus and binds TCF/Lef1 target DNA elements to activate transcription (7, 8). Other experiments suggest that the adenomatous polyposis coli (APC) tumor suppressor gene also plays an important role in Wnt signaling by regulating β-catenin levels (9). APC is phosphorylated by GSK-3β, binds to β-catenin, and facilitates its degradation. Mutations in either APC or β-catenin have been associated with colon carcinomas and melanomas, suggesting these mutations contribute to the development of these types of cancer, implicating the Wnt pathway in tumorigenesis (1).

Although much has been learned about the Wnt signaling pathway over the past several years, only a few of the transcriptionally activated downstream components activated by Wnt have been characterized. Those that have been described cannot account for all of the diverse functions attributed to Wnt signaling. Among the candidate Wnt target genes are those encoding the nodal-related 3 gene, *Xnr3*, a member of the transforming growth factor (TGF)-β superfamily, and the homeobox genes, *engrailed*, *goosecoid*, *twin* (*Xtwn*), and *siamois* (2). A recent report also identifies c-*myc* as a target gene of the Wnt signaling pathway (10).

To identify additional downstream genes in the Wnt signaling pathway that are relevant to the transformed cell phenotype, we used a PCR-based cDNA subtraction strategy, suppression subtractive hybridization (SSH) (11), using RNA isolated from C57MG mouse mammary epithelial cells and C57MG cells stably transformed by a Wnt-1 retrovirus. Overexpression of Wnt-1 in this cell line is sufficient to induce a partially transformed phenotype, characterized by elongated and refractile cells that lose contact inhibition and form a multilayered array (12, 13). We reasoned that genes differentially expressed between these two cell lines might contribute to the transformed phenotype.

In this paper, we describe the cloning and characterization of two genes up-regulated in Wnt-1 transformed cells, *WISP-1* and *WISP-2*, and a third related gene, *WISP-3*. The *WISP* genes are members of the CCN family of growth factors, which includes connective tissue growth factor (CTGF), Cyr61, and *nov*, a family not previously linked to Wnt signaling.

MATERIALS AND METHODS

SSH. SSH was performed by using the PCR-Select cDNA Subtraction Kit (CLONTECH). Tester double-stranded

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: TGF, transforming growth factor; CTGF, connective tissue growth factor; SSH, suppression subtractive hybridization; VWC, von Willebrand factor type C module.

Data deposition: The sequences reported in this paper have been deposited in the Genbank database (accession nos. AF100777, AF100778, AF100779, AF100780, and AF100781).

†To whom reprint requests should be addressed. e-mail: diane@gene.com.

cDNA was synthesized from 2 μg of poly(A)$^+$ RNA isolated from the C57MG/Wnt-1 cell line and driver cDNA from 2 μg of poly(A)$^+$ RNA from the parent C57MG cells. The subtracted cDNA library was subcloned into a pGEM-T vector for further analysis.

cDNA Library Screening. Clones encoding full-length mouse *WISP-1* were isolated by screening a λgt10 mouse embryo cDNA library (CLONTECH) with a 70-bp probe from the original partial clone 568 sequence corresponding to amino acids 128–169. Clones encoding full-length human *WISP-1* were isolated by screening λgt10 lung and fetal kidney cDNA libraries with the same probe at low stringency. Clones encoding full-length mouse and human *WISP-2* were isolated by screening a C57MG/Wnt-1 or human fetal lung cDNA library with a probe corresponding to nucleotides 1463–1512. Full-length cDNAs encoding *WISP-3* were cloned from human bone marrow and fetal kidney libraries.

Expression of Human *WISP* RNA. PCR amplification of first-strand cDNA was performed with human Multiple Tissue cDNA panels (CLONTECH) and 300 μM of each dNTP at 94°C for 1 sec, 62°C for 30 sec, 72°C for 1 min, for 22–32 cycles. *WISP* and glyceraldehyde-3-phosphate dehydrogenase primer sequences are available on request.

***In Situ* Hybridization.** ^{33}P-labeled sense and antisense riboprobes were transcribed from an 897-bp PCR product corresponding to nucleotides 601–1440 of mouse *WISP-1* or a 294-bp PCR product corresponding to nucleotides 82–375 of mouse *WISP-2*. All tissues were processed as described (40).

Radiation Hybrid Mapping. Genomic DNA from each hybrid in the Stanford G3 and Genebridge4 Radiation Hybrid Panels (Research Genetics, Huntsville, AL) and human and hamster control DNAs were PCR-amplified, and the results were submitted to the Stanford or Massachusetts Institute of Technology web servers.

Cell Lines, Tumors, and Mucosa Specimens. Tissue specimens were obtained from the Department of Pathology (University of Pittsburgh) for patients undergoing colon resection and from the University of Leeds, United Kingdom. Genomic DNA was isolated (Qiagen) from the pooled blood of 10 normal human donors, surgical specimens, and the following ATCC human cell lines: SW480, COLO 320DM, HT-29, WiDr, and SW403 (colon adenocarcinomas), SW620 (lymph node metastasis, colon adenocarcinoma), HCT 116 (colon carcinoma), SK-CO-1 (colon adenocarcinoma, ascites), and HM7 (a variant of ATCC colon adenocarcinoma cell line LS 174T). DNA concentration was determined by using Hoechst dye 33258 intercalation fluorimetry. Total RNA was prepared by homogenization in 7 M GuSCN followed by centrifugation over CsCl cushions or prepared by using RNAzol.

Gene Amplification and RNA Expression Analysis. Relative gene amplification and RNA expression of *WISP*s and c-*myc* in the cell lines, colorectal tumors, and normal mucosa were determined by quantitative PCR. Gene-specific primers and fluorogenic probes (sequences available on request) were designed and used to amplify and quantitate the genes. The relative gene copy number was derived by using the formula $2^{(\Delta ct)}$ where ΔCt represents the difference in amplification cycles required to detect the *WISP* genes in peripheral blood lymphocyte DNA compared with colon tumor DNA or colon tumor RNA compared with normal mucosal RNA. The ∂-method was used for calculation of the SE of the gene copy number or RNA expression level. The *WISP*-specific signal was normalized to that of the glyceraldehyde-3-phosphate dehydrogenase housekeeping gene. All TaqMan assay reagents were obtained from Perkin–Elmer Applied Biosystems.

RESULTS

Isolation of *WISP-1* and *WISP-2* by SSH. To identify Wnt-1-inducible genes, we used the technique of SSH using the mouse mammary epithelial cell line C57MG and C57MG cells that stably express Wnt-1 (11). Candidate differentially expressed cDNAs (1,384 total) were sequenced. Thirty-nine percent of the sequences matched known genes or homologues, 32% matched expressed sequence tags, and 29% had no match. To confirm that the transcript was differentially expressed, semiquantitative reverse transcription–PCR and Northern analysis were performed by using mRNA from the C57MG and C57MG/Wnt-1 cells.

Two of the cDNAs, *WISP-1* and *WISP-2*, were differentially expressed, being induced in the C57MG/Wnt-1 cell line, but not in the parent C57MG cells or C57MG cells overexpressing Wnt-4 (Fig. 1 *A* and *B*). Wnt-4, unlike Wnt-1, does not induce the morphological transformation of C57MG cells and has no effect on β-catenin levels (13, 14). Expression of *WISP-1* was up-regulated approximately 3-fold in the C57MG/Wnt-1 cell line and *WISP-2* by approximately 5-fold by both Northern analysis and reverse transcription–PCR.

An independent, but similar, system was used to examine *WISP* expression after Wnt-1 induction. C57MG cells expressing the *Wnt-1* gene under the control of a tetracycline-repressible promoter produce low amounts of Wnt-1 in the repressed state but show a strong induction of *Wnt-1* mRNA and protein within 24 hr after tetracycline removal (8). The levels of Wnt-1 and *WISP* RNA isolated from these cells at various times after tetracycline removal were assessed by quantitative PCR. Strong induction of Wnt-1 mRNA was seen as early as 10 hr after tetracycline removal. Induction of *WISP* mRNA (2- to 6-fold) was seen at 48 and 72 hr (data not shown). These data support our previous observations that show that *WISP* induction is correlated with Wnt-1 expression. Because the induction is slow, occurring after approximately 48 hr, the induction of *WISP*s may be an indirect response to Wnt-1 signaling.

cDNA clones of human *WISP-1* were isolated and the sequence compared with mouse *WISP-1*. The cDNA sequences of mouse and human *WISP-1* were 1,766 and 2,830 bp in length, respectively, and encode proteins of 367 aa, with predicted relative molecular masses of ≈40,000 (M_r 40 K). Both have hydrophobic N-terminal signal sequences, 38 conserved cysteine residues, and four potential N-linked glycosylation sites and are 84% identical (Fig. 2*A*).

Full-length cDNA clones of mouse and human *WISP-2* were 1,734 and 1,293 bp in length, respectively, and encode proteins of 251 and 250 aa, respectively, with predicted relative molecular masses of ≈27,000 (M_r 27 K) (Fig. 2*B*). Mouse and human *WISP-2* are 73% identical. Human *WISP-2* has no potential N-linked glycosylation sites, and mouse *WISP-2* has one at

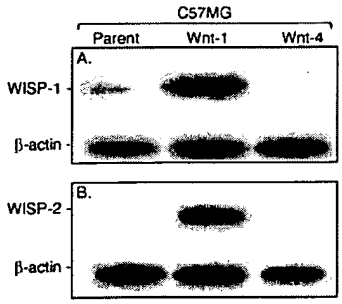


FIG. 1. *WISP-1* and *WISP-2* are induced by Wnt-1, but not Wnt-4, expression in C57MG cells. Northern analysis of *WISP-1* (*A*) and *WISP-2* (*B*) expression in C57MG, C57MG/Wnt-1, and C57MG/Wnt-4 cells. Poly(A)$^+$ RNA (2 μg) was subjected to Northern blot analysis and hybridized with a 70-bp mouse *WISP-1*-specific probe (amino acids 278–300) or a 190-bp *WISP-2*-specific probe (nucleotides 1438–1627) in the 3′ untranslated region. Blots were rehybridized with human β-actin probe.

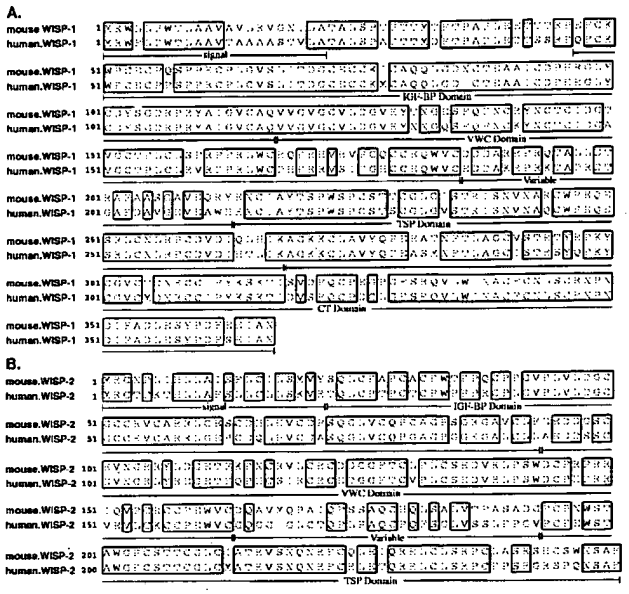


FIG. 2. Encoded amino acid sequence alignment of mouse and human *WISP-1* (*A*) and mouse and human *WISP-2* (*B*). The potential signal sequence, insulin-like growth factor-binding protein (IGF-BP), VWC, thrombospondin (TSP), and C-terminal (CT) domains are underlined.

position 197. *WISP-2* has 28 cysteine residues that are conserved among the 38 cysteines found in *WISP-1*.

Identification of *WISP-3*. To search for related proteins, we screened expressed sequence tag (EST) databases with the WISP-1 protein sequence and identified several ESTs as potentially related sequences. We identified a homologous protein that we have called WISP-3. A full-length human *WISP-3* cDNA of 1,371 bp was isolated corresponding to those ESTs that encode a 354-aa protein with a predicted molecular mass of 39,293. WISP-3 has two potential N-linked glycosylation sites and 36 cysteine residues. An alignment of the three human WISP proteins shows that WISP-1 and WISP-3 are the most similar (42% identity), whereas WISP-2 has 37% identity with WISP-1 and 32% identity with WISP-3 (Fig. 3*A*).

***WISPs* Are Homologous to the CTGF Family of Proteins.** Human *WISP-1*, *WISP-2*, and *WISP-3* are novel sequences; however, mouse *WISP-1* is the same as the recently identified *Elm1* gene. *Elm1* is expressed in low, but not high, metastatic mouse melanoma cells, and suppresses the *in vivo* growth and metastatic potential of K-1735 mouse melanoma cells (15). Human and mouse *WISP-2* are homologous to the recently described rat gene, *rCop-1* (16). Significant homology (36–44%) was seen to the CCN family of growth factors. This family includes three members, CTGF, Cyr61, and the protooncogene *nov*. CTGF is a chemotactic and mitogenic factor for fibroblasts that is implicated in wound healing and fibrotic disorders and is induced by TGF-β (17). Cyr61 is an extracellular matrix signaling molecule that promotes cell adhesion, proliferation, migration, angiogenesis, and tumor growth (18, 19). *nov* (nephroblastoma overexpressed) is an immediate early gene associated with quiescence and found altered in Wilms tumors (20). The proteins of the CCN family share functional, but not sequence, similarity to Wnt-1. All are secreted, cysteine-rich heparin binding glycoproteins that associate with the cell surface and extracellular matrix.

WISP proteins exhibit the modular architecture of the CCN family, characterized by four conserved cysteine-rich domains (Fig. 3*B*) (21). The N-terminal domain, which includes the first 12 cysteine residues, contains a consensus sequence (GCGCCXXC) conserved in most insulin-like growth factor (IGF)-binding proteins (BP). This sequence is conserved in WISP-2 and WISP-3, whereas WISP-1 has a glutamine in the third position instead of a glycine. CTGF recently has been shown to specifically bind IGF (22) and a truncated *nov* protein lacking the IGF-BP domain is oncogenic (23). The von Willebrand factor type C module (VWC), also found in certain collagens and mucins, covers the next 10 cysteine residues, and is thought to participate in protein complex formation and oligomerization (24). The VWC domain of WISP-3 differs from all CCN family members described previously, in that it contains only six of the 10 cysteine residues (Fig. 3 *A* and *B*). A short variable region follows the VWC domain. The third module, the thrombospondin (TSP) domain is involved in binding to sulfated glycoconjugates and contains six cysteine residues and a conserved WSxCSxxCG motif first identified in thrombospondin (25). The C-terminal (CT) module containing the remaining 10 cysteines is thought to be involved in dimerization and receptor binding (26). The CT domain is present in all CCN family members described to date but is absent in WISP-2 (Fig. 3 *A* and *B*). The existence of a putative signal sequence and the absence of a transmembrane domain suggest that WISPs are secreted proteins, an observation supported by an analysis of their expression and secretion from mammalian cell and baculovirus cultures (data not shown).

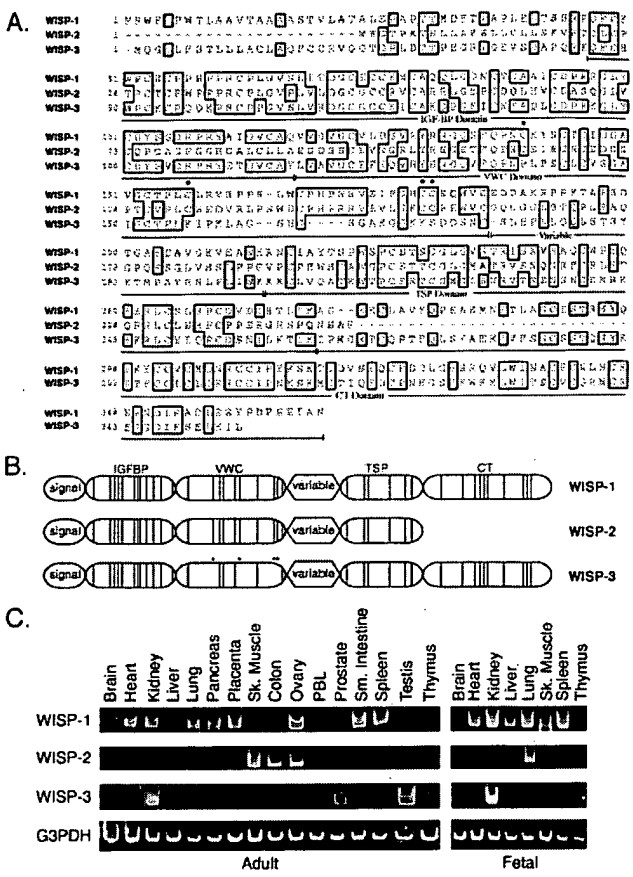


FIG. 3. (*A*) Encoded amino acid sequence alignment of human WISPs. The cysteine residues of WISP-1 and WISP-2 that are not present in WISP-3 are indicated with a dot. (*B*) Schematic representation of the WISP proteins showing the domain structure and cysteine residues (vertical lines). The four cysteine residues in the VWC domain that are absent in WISP-3 are indicated with a dot. (*C*) Expression of *WISP* mRNA in human tissues. PCR was performed on human multiple-tissue cDNA panels (CLONTECH) from the indicated adult and fetal tissues.

Expression of *WISP* mRNA in Human Tissues. Tissue-specific expression of human *WISP*s was characterized by PCR

analysis on adult and fetal multiple tissue cDNA panels. *WISP-1* expression was seen in the adult heart, kidney, lung, pancreas, placenta, ovary, small intestine, and spleen (Fig. 3*C*). Little or no expression was detected in the brain, liver, skeletal muscle, colon, peripheral blood leukocytes, prostate, testis, or thymus. *WISP-2* had a more restricted tissue expression and was detected in adult skeletal muscle, colon, ovary, and fetal lung. Predominant expression of *WISP-3* was seen in adult kidney and testis and fetal kidney. Lower levels of *WISP-3* expression were detected in placenta, ovary, prostate, and small intestine.

***In Situ* Localization of *WISP-1* and *WISP-2*.** Expression of *WISP-1* and *WISP-2* was assessed by *in situ* hybridization in mammary tumors from Wnt-1 transgenic mice. Strong expression of *WISP-1* was observed in stromal fibroblasts lying within the fibrovascular tumor stroma (Fig. 4 *A–D*). However, low-level *WISP-1* expression also was observed focally within tumor cells (data not shown). No expression was observed in normal breast. Like *WISP-1*, *WISP-2* expression also was seen in the tumor stroma in breast tumors from Wnt-1 transgenic animals (Fig. 4 *E–H*). However, *WISP-2* expression in the stroma was in spindle-shaped cells adjacent to capillary vessels, whereas the predominant cell type expressing *WISP-1* was the stromal fibroblasts.

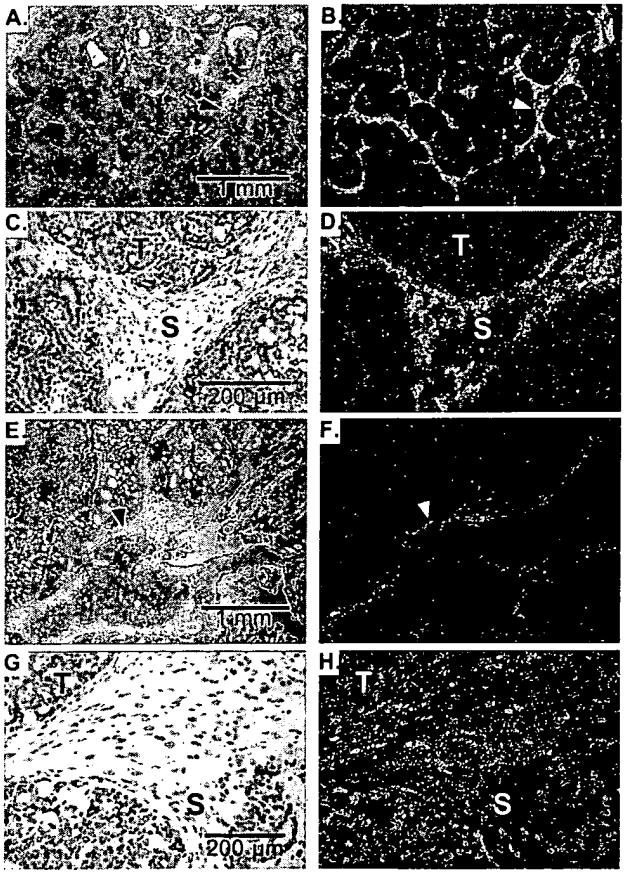


FIG. 4. (*A*, *C*, *E*, and *G*) Representative hematoxylin/eosin-stained images from breast tumors in Wnt-1 transgenic mice. The corresponding dark-field images showing *WISP-1* expression are shown in *B* and *D*. The tumor is a moderately well-differentiated adenocarcinoma showing evidence of adenoid cystic change. At low power (*A* and *B*), expression of *WISP-1* is seen in the delicate branching fibrovascular tumor stroma (arrowhead). At higher magnification, expression is seen in the stromal(s) fibroblasts (*C* and *D*), and tumor cells are negative. Focal expression of *WISP-1*, however, was observed in tumor cells in some areas. Images of *WISP-2* expression are shown in *E–H*. At low power (*E* and *F*), expression of WISP-2 is seen in cells lying within the fibrovascular tumor stroma. At higher magnification, these cells appeared to be adjacent to capillary vessels whereas tumor cells are negative (*G* and *H*).

Chromosome Localization of the *WISP* Genes. The chromosomal location of the human *WISP* genes was determined by radiation hybrid mapping panels. *WISP-1* is approximately 3.48 cR from the meiotic marker AFM259xc5 [logarithm of odds (lod) score 16.31] on chromosome 8q24.1 to 8q24.3, in the same region as the human locus of the *nov*H family member (27) and roughly 4 Mbs distal to c-*myc* (28). Preliminary fine mapping indicates that *WISP-1* is located near D8S1712 STS. *WISP-2* is linked to the marker SHGC-33922 (lod = 1,000) on chromosome 20q12–20q13.1. Human *WISP-3* mapped to chromosome 6q22–6q23 and is linked to the marker AFM211ze5 (lod = 1,000). *WISP-3* is approximately 18 Mbs proximal to CTGF and 23 Mbs proximal to the human cellular oncogene *MYB* (27, 29).

Amplification and Aberrant Expression of *WISPs* in Human Colon Tumors. Amplification of protooncogenes is seen in many human tumors and has etiological and prognostic significance. For example, in a variety of tumor types, c-*myc* amplification has been associated with malignant progression and poor prognosis (30). Because *WISP-1* resides in the same general chromosomal location (8q24) as c-*myc*, we asked whether it was a target of gene amplification, and, if so, whether this amplification was independent of the c-*myc* locus. Genomic DNA from human colon cancer cell lines was assessed by quantitative PCR and Southern blot analysis. (Fig. 5 *A* and *B*). Both methods detected similar degrees of *WISP-1* amplification. Most cell lines showed significant (2- to 4-fold) amplification, with the HT-29 and WiDr cell lines demonstrating an 8-fold increase. Significantly, the pattern of amplification observed did not correlate with that observed for c-*myc*, indicating that the c-*myc* gene is not part of the amplicon that involves the *WISP-1* locus.

We next examined whether the *WISP* genes were amplified in a panel of 25 primary human colon adenocarcinomas. The relative *WISP* gene copy number in each colon tumor DNA was compared with pooled normal DNA from 10 donors by quantitative PCR (Fig. 6). The copy number of *WISP-1* and *WISP-2* was significantly greater than one, approximately 2-fold for *WISP-1* in about 60% of the tumors and 2- to 4-fold for *WISP-2* in 92% of the tumors ($P < 0.001$ for each). The copy number for *WISP-3* was indistinguishable from one ($P = 0.166$). In addition, the copy number of *WISP-2* was significantly higher than that of *WISP-1* ($P < 0.001$).

The levels of *WISP* transcripts in RNA isolated from 19 adenocarcinomas and their matched normal mucosa were

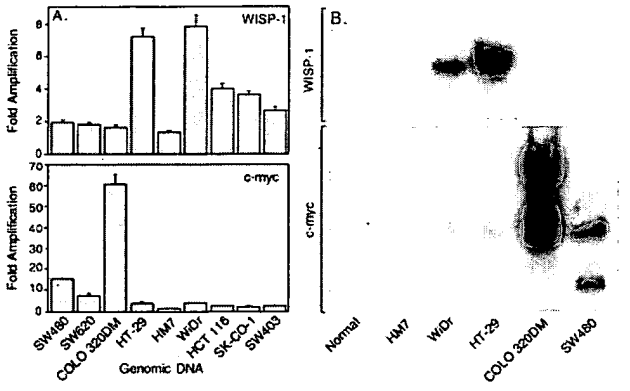


FIG. 5. Amplification of *WISP-1* genomic DNA in colon cancer cell lines. (*A*) Amplification in cell line DNA was determined by quantitative PCR. (*B*) Southern blots containing genomic DNA (10 μg) digested with *Eco*RI (*WISP-1*) or *Xba*I (c-*myc*) were hybridized with a 100-bp human *WISP-1* probe (amino acids 186–219) or a human c-*myc* probe (located at bp 1901–2000). The *WISP* and *myc* genes are detected in normal human genomic DNA after a longer film exposure.

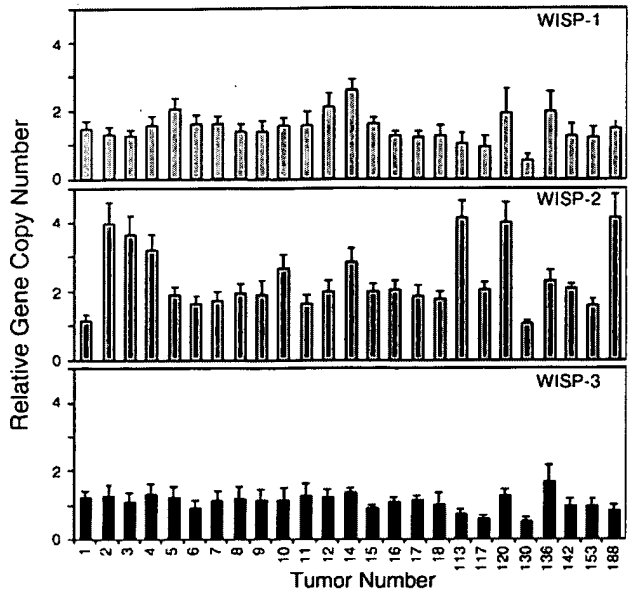


FIG. 6. Genomic amplification of *WISP* genes in human colon tumors. The relative gene copy number of the *WISP* genes in 25 adenocarcinomas was assayed by quantitative PCR, by comparing DNA from primary human tumors with pooled DNA from 10 healthy donors. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least three times.

assessed by quantitative PCR (Fig. 7). The level of *WISP-1* RNA present in tumor tissue varied but was significantly increased (2- to >25-fold) in 84% (16/19) of the human colon tumors examined compared with normal adjacent mucosa. Four of 19 tumors showed greater than 10-fold overexpression. In contrast, in 79% (15/19) of the tumors examined, *WISP-2* RNA expression was significantly lower in the tumor than the mucosa. Similar to *WISP-1*, *WISP-3* RNA was overexpressed in 63% (12/19) of the colon tumors compared with the normal

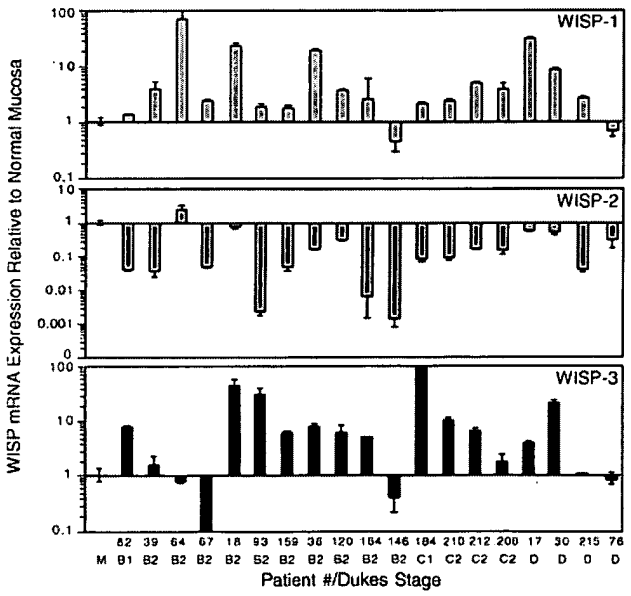


FIG. 7. *WISP* RNA expression in primary human colon tumors relative to expression in normal mucosa from the same patient. Expression of *WISP* mRNA in 19 adenocarcinomas was assayed by quantitative PCR. The Dukes stage of the tumor is listed under the sample number. The data are means ± SEM from one experiment done in triplicate. The experiment was repeated at least twice.

mucosa. The amount of overexpression of *WISP-3* ranged from 4- to >40-fold.

DISCUSSION

One approach to understanding the molecular basis of cancer is to identify differences in gene expression between cancer cells and normal cells. Strategies based on assumptions that steady-state mRNA levels will differ between normal and malignant cells have been used to clone differentially expressed genes (31). We have used a PCR-based selection strategy, SSH, to identify genes selectively expressed in C57MG mouse mammary epithelial cells transformed by Wnt-1.

Three of the genes isolated, *WISP-1*, *WISP-2*, and *WISP-3*, are members of the CCN family of growth factors, which includes CTGF, Cyr61, and *nov*, a family not previously linked to Wnt signaling.

Two independent experimental systems demonstrated that *WISP* induction was associated with the expression of Wnt-1. The first was C57MG cells infected with a Wnt-1 retroviral vector or C57MG cells expressing Wnt-1 under the control of a tetracyline-repressible promoter, and the second was in Wnt-1 transgenic mice, where breast tissue expresses Wnt-1, whereas normal breast tissue does not. No *WISP* RNA expression was detected in mammary tumors induced by polyoma virus middle T antigen (data not shown). These data suggest a link between Wnt-1 and *WISP*s in that in these two situations, *WISP* induction was correlated with Wnt-1 expression.

It is not clear whether the *WISP*s are directly or indirectly induced by the downstream components of the Wnt-1 signaling pathway (i.e., β-catenin-TCF-1/Lef1). The increased levels of *WISP* RNA were measured in Wnt-1-transformed cells, hours or days after Wnt-1 transformation. Thus, *WISP* expression could result from Wnt-1 signaling directly through β-catenin transcription factor regulation or alternatively through Wnt-1 signaling turning on a transcription factor, which in turn regulates *WISP*s.

The *WISP*s define an additional subfamily of the CCN family of growth factors. One striking difference observed in the protein sequence of WISP-2 is the absence of a CT domain, which is present in CTGF, Cyr61, *nov*, WISP-1, and WISP-3. This domain is thought to be involved in receptor binding and dimerization. Growth factors, such as TGF-β, platelet-derived growth factor, and nerve growth factor, which contain a cystine knot motif exist as dimers (32). It is tempting to speculate that WISP-1 and WISP-3 may exist as dimers, whereas WISP-2 exists as a monomer. If the CT domain is also important for receptor binding, WISP-2 may bind its receptor through a different region of the molecule than the other CCN family members. No specific receptors have been identified for CTGF or *nov*. A recent report has shown that integrin $\alpha_v\beta_3$ serves as an adhesion receptor for Cyr61 (33).

The strong expression of *WISP-1* and *WISP-2* in cells lying within the fibrovascular tumor stroma in breast tumors from Wnt-1 transgenic animals is consistent with previous observations that transcripts for the related CTGF gene are primarily expressed in the fibrous stroma of mammary tumors (34). Epithelial cells are thought to control the proliferation of connective tissue stroma in mammary tumors by a cascade of growth factor signals similar to that controlling connective tissue formation during wound repair. It has been proposed that mammary tumor cells or inflammatory cells at the tumor interstitial interface secrete TGF-β1, which is the stimulus for stromal proliferation (34). TGF-β1 is secreted by a large percentage of malignant breast tumors and may be one of the growth factors that stimulates the production of CTGF and WISPs in the stroma.

It was of interest that *WISP-1* and *WISP-2* expression was observed in the stromal cells that surrounded the tumor cells

(epithelial cells) in the Wnt-1 transgenic mouse sections of breast tissue. This finding suggests that paracrine signaling could occur in which the stromal cells could supply WISP-1 and WISP-2 to regulate tumor cell growth on the WISP extracellular matrix. Stromal cell-derived factors in the extracellular matrix have been postulated to play a role in tumor cell migration and proliferation (35). The localization of *WISP-1* and *WISP-2* in the stromal cells of breast tumors supports this paracrine model.

An analysis of *WISP-1* gene amplification and expression in human colon tumors showed a correlation between DNA amplification and overexpression, whereas overexpression of *WISP-3* RNA was seen in the absence of DNA amplification. In contrast, *WISP-2* DNA was amplified in the colon tumors, but its mRNA expression was significantly reduced in the majority of tumors compared with the expression in normal colonic mucosa from the same patient. The gene for human *WISP-2* was localized to chromosome 20q12–20q13, at a region frequently amplified and associated with poor prognosis in node negative breast cancer and many colon cancers, suggesting the existence of one or more oncogenes at this locus (36–38). Because the center of the 20q13 amplicon has not yet been identified, it is possible that the apparent amplification observed for *WISP-2* may be caused by another gene in this amplicon.

A recent manuscript on *rCop-1*, the rat orthologue of *WISP-2*, describes the loss of expression of this gene after cell transformation, suggesting it may be a negative regulator of growth in cell lines (16). Although the mechanism by which *WISP-2* RNA expression is down-regulated during malignant transformation is unknown, the reduced expression of *WISP-2* in colon tumors and cell lines suggests that it may function as a tumor suppressor. These results show that the *WISP* genes are aberrantly expressed in colon cancer and suggest that their altered expression may confer selective growth advantage to the tumor.

Members of the Wnt signaling pathway have been implicated in the pathogenesis of colon cancer, breast cancer, and melanoma, including the tumor suppressor gene adenomatous polyposis coli and β-catenin (39). Mutations in specific regions of either gene can cause the stabilization and accumulation of cytoplasmic β-catenin, which presumably contributes to human carcinogenesis through the activation of target genes such as the *WISP*s. Although the mechanism by which Wnt-1 transforms cells and induces tumorigenesis is unknown, the identification of *WISP*s as genes that may be regulated downstream of Wnt-1 in C57MG cells suggests they could be important mediators of Wnt-1 transformation. The amplification and altered expression patterns of the *WISP*s in human colon tumors may indicate an important role for these genes in tumor development.

We thank the DNA synthesis group for oligonucleotide synthesis, T. Baker for technical assistance, P. Dowd for radiation hybrid mapping, K. Willert and R. Nusse for the tet-repressible C57MG/Wnt-1 cells, V. Dixit for discussions, and D. Wood and A. Bruce for artwork.

1. Cadigan, K. M. & Nusse, R. (1997) *Genes Dev.* **11,** 3286–3305.
2. Dale, T. C. (1998) *Biochem. J.* **329,** 209–223.
3. Nusse, R. & Varmus, H. E. (1982) *Cell* **31,** 99–109.
4. van Ooyen, A. & Nusse, R. (1984) *Cell* **39,** 233–240.
5. Tsukamoto, A. S., Grosschedl, R., Guzman, R. C., Parslow, T. & Varmus, H. E. (1988) *Cell* **55,** 619–625.
6. Brown, J. D. & Moon, R. T. (1998) *Curr. Opin. Cell. Biol.* **10,** 182–187.
7. Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) *Cell* **86,** 391–399.
8. Korinek, V., Barker, N., Willert, K., Molenaar, M., Roose, J., Wagenaar, G., Markman, M., Lamers, W., Destree, O. & Clevers, H. (1998) *Mol. Cell. Biol.* **18,** 1248–1256.
9. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3046–3050.
10. He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* **281,** 1509–1512.
11. Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. & Siebert, P. D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 6025–6030.
12. Brown, A. M., Wildin, R. S., Prendergast, T. J. & Varmus, H. E. (1986) *Cell* **46,** 1001–1009.
13. Wong, G. T., Gavin, B. J. & McMahon, A. P. (1994) *Mol. Cell. Biol.* **14,** 6278–6286.
14. Shimizu, H., Julius, M. A., Giarre, M., Zheng, Z., Brown, A. M. & Kitajewski, J. (1997) *Cell Growth Differ.* **8,** 1349–1358.
15. Hashimoto, Y., Shindo-Okada, N., Tani, M., Nagamachi, Y., Takeuchi, K., Shiroishi, T., Toma, H. & Yokota, J. (1998) *J. Exp. Med.* **187,** 289–296.
16. Zhang, R., Averboukh, L., Zhu, W., Zhang, H., Jo, H., Dempsey, P. J., Coffey, R. J., Pardee, A. B. & Liang, P. (1998) *Mol. Cell. Biol.* **18,** 6131–6141.
17. Grotendorst, G. R. (1997) *Cytokine Growth Factor Rev.* **8,** 171–179.
18. Kireeva, M. L., Mo, F. E., Yang, G. P. & Lau, L. F. (1996) *Mol. Cell. Biol.* **16,** 1326–1334.
19. Babic, A. M., Kireeva, M. L., Kolesnikova, T. V. & Lau, L. F. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6355–6360.
20. Martinerie, C., Huff, V., Joubert, I., Badzioch, M., Saunders, G., Strong, L. & Perbal, B. (1994) *Oncogene* **9,** 2729–2732.
21. Bork, P. (1993) *FEBS Lett.* **327,** 125–130.
22. Kim, H. S., Nagalla, S. R., Oh, Y., Wilson, E., Roberts, C. T., Jr. & Rosenfeld, R. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 12981–12986.
23. Joliot, V., Martinerie, C., Dambrine, G., Plassiart, G., Brisac, M., Crochet, J. & Perbal, B. (1992) *Mol. Cell. Biol.* **12,** 10–21.
24. Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G. & Sadler, J. E. (1989) *J. Biol. Chem.* **264,** 19514–19527.
25. Holt, G. D., Pangburn, M. K. & Ginsburg, V. (1990) *J. Biol. Chem.* **265,** 2852–2855.
26. Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A. & Pannekoek, H. (1991) *J. Cell. Biol.* **113,** 195–205.
27. Martinerie, C., Viegas-Pequignot, E., Guenard, I., Dutrillaux, B., Nguyen, V. C., Bernheim, A. & Perbal, B. (1992) *Oncogene* **7,** 2529–2534.
28. Takahashi, E., Hori, T., O'Connell, P., Leppert, M. & White, R. (1991) *Cytogenet. Cell. Genet.* **57,** 109–111.
29. Meese, E., Meltzer, P. S., Witkowski, C. M. & Trent, J. M. (1989) *Genes Chromosomes Cancer* **1,** 88–94.
30. Garte, S. J. (1993) *Crit. Rev. Oncog.* **4,** 435–449.
31. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276,** 1268–1272.
32. Sun, P. D. & Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* **24,** 269–291.
33. Kireeva, M. L., Lam, S. C. T. & Lau, L. F. (1998) *J. Biol. Chem.* **273,** 3090–3096.
34. Frazier, K. S. & Grotendorst, G. R. (1997) *Int. J. Biochem. Cell. Biol.* **29,** 153–161.
35. Wernert, N. (1997) *Virchows Arch.* **430,** 433–443.
36. Tanner, M. M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W. L., *et al.* (1994) *Cancer Res.* **54,** 4257–4260.
37. Brinkmann, U., Gallo, M., Polymeropoulos, M. H. & Pastan, I. (1996) *Genome Res.* **6,** 187–194.
38. Bischoff, J. R., Anderson, L., Zhu, Y., Mossie, K., Ng, L., Souza, B., Schryver, B., Flanagan, P., Clairvoyant, F., Ginther, C., *et al.* (1998) *EMBO J.* **17,** 3052–3065.
39. Morin, P. J., Sparks, A. B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **275,** 1787–1790.
40. Lu, L. H. & Gillett, N. (1994) *Cell Vision* **1,** 169–176.

Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 4049–4052, June 1986
Medical Sciences

Variable expression of the translocated c-*abl* oncogene in Philadelphia-chromosome-positive B-lymphoid cell lines from chronic myelogenous leukemia patients

JAMES B. KONOPKA*‡, STEVEN CLARK*, JAMI MCLAUGHLIN*, MASAKUZU NITTA†, YOSHIRO KATO†, ANNABEL STRIFE†, BAYARD CLARKSON†, AND OWEN N. WITTE*§

*Department of Microbiology and Molecular Biology Institute, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90024; and †The Laboratory of Hematopoietic Cell Kinetics and The Laboratory of Cancer Genetics and Cytogenetics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021

Communicated by Michael Potter, February 10, 1986

ABSTRACT **The consistent cytogenetic translocation of chronic myelogenous leukemia (the Philadelphia chromosome, Ph^1) has been observed in cells of multiple hematopoietic lineages. This translocation creates a chimeric gene composed of breakpoint-cluster-region (*bcr*) sequences from chromosome 22 fused to a portion of the *abl* oncogene on chromosome 9. The resulting gene product ($P210^{c\text{-}abl}$) resembles the transforming protein of the Abelson murine leukemia virus in its structure and tyrosine kinase activity. $P210^{c\text{-}abl}$ is expressed in Ph^1-positive cell lines of myeloid lineage and in clinical specimens with myeloid predominance. We show here that Epstein–Barr virus-transformed B-lymphocyte lines that retain Ph^1 can express $P210^{c\text{-}abl}$. The level of expression in these B-cell lines is generally lower and more variable than that observed for myeloid lines. Protein expression is not related to amplification of the *abl* gene but to variation in the level of *bcr-abl* mRNA produced from a single Ph^1 template.**

Chronic myelogenous leukemia (CML) is a disease of the pluripotent stem cell (1). In greater than 95% of patients, the leukemic cells contain the cytogenetic marker known as the Philadelphia chromosome, or Ph^1 (2). This reciprocal translocation event between the long arms of chromosomes 9 and 22 has been used as a disease-specific marker for diagnosis and evaluation of therapy. Multiple hematopoietic lineages, including myeloid and B-lymphoid, contain Ph^1 in early or chronic phase, as well as in the more acute accelerated and blast crisis phases of the disease.

One molecular consequence of Ph^1 is the translocation of the chromosomal arm containing the c-*abl* gene on chromosome 9 into the middle of the breakpoint-cluster region (*bcr*) gene on chromosome 22 (3–6). Although the precise translocation breakpoints are variable, an RNA-splicing mechanism generates a very similar 8-kilobase (kb) mRNA in each case (5–9). The hybrid *bcr-abl* message encodes a structurally altered form of the *abl* oncogene product, called $P210^{c\text{-}abl}$ (10–13), with an amino-terminal segment derived from a portion of the exons of *bcr* on chromosome 22 and a carboxyl-terminal segment derived from a major portion of the exons of the c-*abl* gene on chromosome 9. The chimeric structure of *bcr-abl* and the resulting $P210^{c\text{-}abl}$ is similar to the structure of the Abelson murine leukemia virus *gag-abl* genome and resulting $P160^{v\text{-}abl}$ transforming gene product. Both proteins have very similar tyrosine kinase activities (10, 11, 14) which can be distinguished by their relative stability to denaturing detergents and by their ATP requirements from the recently described tyrosine kinase activity of the c-*abl* gene product (15).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

In concert with structural modification of the amino-terminal portion of the *abl* gene, increased level of expression has been implicated in activation of c-*abl* oncogenic potential. Myeloid and erythroid cell lines and clinical samples derived from acute-phase CML patients contain about 10-fold higher levels of the 8-kb *bcr-abl* mRNA and $P210^{c\text{-}abl}$ than the c-*abl* mRNA forms (6 and 7 kb) and $P145^{c\text{-}abl}$ gene product (5, 8, 9, 11). The higher level of expression of the chimeric *bcr-abl* message in acute-phase cells is not likely to be solely due to the presence of the *bcr* promoter sequences at the 5′ end of the gene, since the normal 4.5-kb and 6.7-kb *bcr*-encoded mRNA species are expressed at an even lower level than the normal c-*abl* messages (5, 6).

We have analyzed a series of Epstein–Barr virus-immortalized B-lymphoid cell lines derived from CML patients (16). With such *in vitro* clonal cell lines, we can evaluate whether the presence of Ph^1 always results in synthesis of the chimeric *bcr-abl* message and protein, and whether the quantitative expression varies for cells of B-lymphoid lineage as compared to previously examined myeloid cell lines. Our results show that cell lines that retain Ph^1 do express *bcr-abl* message and protein, but that the level is generally lower and more variable than previously seen for myeloid cell lines. The demonstration that the Ph^1 chromosomal template can vary in its level of expression of $P210^{c\text{-}abl}$ suggests that secondary mechanisms, beyond the translocation itself, contribute to the regulation of the *bcr-abl* gene in different cell types or subclones that derive from the affected stem cell.

MATERIALS AND METHODS

Cells and Cell Labelings. Epstein–Barr virus-transformed B-lymphoid cell lines were established from peripheral blood samples of chronic- and acute-phase CML patients as reported (16). The cell lines are designated according to patient number, karyotype, and lineage. For example, SK-CML7Bt(9,22)-33 refers to CML patient 7, B-lymphoid cell line, 9;22 translocation (Ph^1), cell line 33; and SK-CML7BN-2 refers to B-cell line 2 with a normal karyotype derived from the same patient. Repeat karyotype analysis was performed to verify the retention of Ph^1 just prior to analysis for *abl* protein and RNA. Cells were maintained in RPMI 1640 medium with 20% fetal bovine serum. We have not observed any consistent pattern of *in vitro* growth rate that correlates to the stage of disease at the time of transformation with Epstein–Barr virus. Cells (1.5×10^7) were washed twice with Dulbecco's modified Eagle's medium lacking phosphate and

Abbreviations: *bcr*, breakpoint-cluster region; CML, chronic myelogenous leukemia; kb, kilobase(s).

‡Present address: Department of Genetics, University of Washington, Seattle, WA 98195.
§To whom correspondence should be addressed.

supplemented with 5% dialyzed fetal bovine serum. Cells were then resuspended in 2 ml of the minimal medium. Labeling was started with the addition of [^{32}P]orthophosphate (1 mCi/ml; ICN; 1 Ci = 37 GBq) and continued at 37°C for 3–4 hr.

Immunoprecipitation and Immunoblotting. Immunoprecipitations were carried out as described (10). Cells (1.5×10^7) were washed with phosphate-buffered saline and extracted with 3–5 ml of phosphate lysis buffer (1% Triton X-100/0.1 $NaDodSO_4$/0.5% deoxycholate/10 mM Na_2HPO_4, pH 7.5/100 mM NaCl) with 5 mM EDTA and 5 mM phenylmethylsulfonyl fluoride. Extracts were clarified by centrifugation and precipitated with normal or rabbit anti-abl sera (anti-pEX-2 or anti-pEX-5) (17). The precipitated proteins were electrophoresed in a $NaDodSO_4$/8% polyacrylamide gel. ^{32}P-labeled proteins were detected by autoradiography. Alternatively, *abl* proteins were detected by immunoblotting. Extracts from unlabeled cells were clarified, and proteins were concentrated by immunoprecipitation with rabbit antisera against *abl*-encoded proteins [anti-pEX-2 and anti-pEX-5 combined (17)] and then fractionated in 8% acrylamide gels. The proteins were transferred from the gel to nitrocellulose filters, using protease-facilitated transfer (18). The *abl*-encoded proteins were detected using murine monoclonal antibodies as a probe and peroxidase-conjugated goat anti-mouse second stage antibody (Bio-Rad) for development. Rabbit antisera and mouse monoclonal antibodies to *abl* proteins were prepared using bacterially expressed regions of the v-*abl* protein as immunogens (17, 19). Anti-pEX-2 antibodies react with the internal tyrosine kinase domain and anti-pEX-5 antibodies react with the carboxyl-terminal segment of the *abl* proteins.

RNA Analysis. RNA was extracted from 10^8 cells by the $NaDodSO_4$/urea/phenol method (20). Polyadenylylated RNA was purified by oligo(dT) affinity chromatography. Samples were electrophoresed in a 1% agarose/formaldehyde gel and transferred to nitrocellulose. *abl* RNA species were detected by hybridization with a nick-translated v-*abl* fragment probe (21).

DNA Analysis. DNA was prepared from 5×10^7 cells of each cell line and processed for Southern blots with a v-*abl* probe as described (21).

RESULTS

Variable Levels of P210$^{c\text{-}abl}$ Are Detected in Ph^1-Positive Cell Lines. Ph^1-positive and Ph^1-negative, Epstein–Barr virus-transformed B-lymphocyte cell lines derived from the same patient were examined for P210$^{c\text{-}abl}$ synthesis by immunoprecipitation of [^{32}P]orthophosphate-labeled cell extracts with anti-abl sera (Fig. 1). The normal c-*abl* protein P145$^{c\text{-}abl}$ was detected at a similar level in multiple Ph^1-positive and Ph^1-negative cell lines. P210$^{c\text{-}abl}$ was only detected in the Ph^1-positive cell lines because the *bcr-abl* chimeric gene which encodes P210$^{c\text{-}abl}$ resides on the Ph^1 (4, 5, 11, 13). The level of P210$^{c\text{-}abl}$ was about 4- to 5-fold higher than the level of P145$^{c\text{-}abl}$ in the SK-CML7Bt-33 cell line (Fig. 1*A*, +). The Ph^1-positive erythroid-progenitor cell line K562 (*C*) showed a level of P210$^{c\text{-}abl}$ about 10-fold higher than P145$^{c\text{-}abl}$. However, the level of P210$^{c\text{-}abl}$ was about one-fifth that of P145$^{c\text{-}abl}$ in the Ph^1-positive SK-CML16Bt-1 cell line (Fig. 1*B*, +). Comparison of different autoradiographic exposures roughly indicated that the level of P210$^{c\text{-}abl}$ varies over a 20-fold range between these Ph^1-positive B-cell lines. Analysis of four additional Ph^1-positive B-cell lines demonstrated that the level of P210$^{c\text{-}abl}$ fell into two general classes; some cell lines had a level of P210$^{c\text{-}abl}$ similar to SK-CML7Bt-33 and others had the low level similar to SK-CML16Bt-1 (Table 1). This differs from previous studies with Ph^1-positive myeloid cell lines and patient samples derived from acute-phase CML patients, in which P210$^{c\text{-}abl}$ was detected at a 10-fold higher level than P145$^{c\text{-}abl}$ (refs. 10 and 11; Table 1). There was no large difference in level of chimeric mRNA and P210$^{c\text{-}abl}$ expressed in four myeloid/erythroid-lineage Ph^1-positive cell lines (K562, EM2, EM3, CML22, and BV173; refs. 9 and 11), despite a 4- to 5-fold amplification of *abl*-related sequences in the K562 cell line.

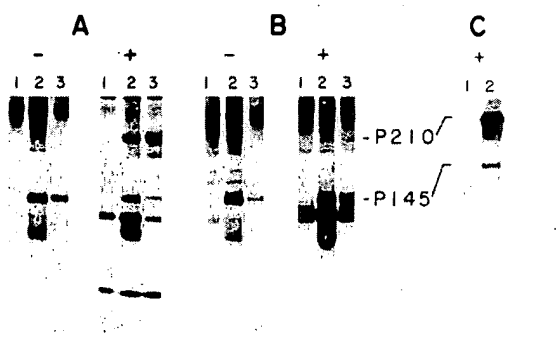


FIG. 1. Detection of variable levels of P210$^{c\text{-}abl}$ in Ph^1-positive B-cell lines. Production of P145$^{c\text{-}abl}$ and P210$^{c\text{-}abl}$ in Epstein–Barr virus-transformed B-cell lines derived from a blast-crisis (*A*) and a chronic-phase (*B*) CML patient was examined by metabolic labeling with [^{32}P]orthophosphate and immunoprecipitation. Ph^1-negative (−) and Ph^1-positive (+) cell lines derived from each patient were analyzed. The Ph^1-negative cell line in *A*,− is SK-CML7BN-2 and in *B*,− is SK-CML16BN-1. The Ph^1-positive cell line in *A*,+ is SK-CML7Bt-33 and in *B*,+ is SK-CML16Bt-1. The K562 cell line, a Ph^1-positive erythroid progenitor cell line spontaneously derived from a blast-crisis patient (33), is represented in *C*. Cells (1.5×10^7) were metabolically labeled with 2 mCi of [^{32}P]orthophosphate for 3–4 hr and then were extracted and clarified by centrifugation. Samples were immunoprecipitated with control normal serum (lanes 1), anti-pEX-2 (lanes 2), or anti-pEX-5 (lanes 3) and analyzed by $NaDodSO_4$/8% PAGE followed by autoradiography with an intensifying screen (3 days for *A* and *C*, 10 days for *B*).

Detection of different levels of P210$^{c\text{-}abl}$ in Fig. 1 could be due to decreased phosphorylation of P210$^{c\text{-}abl}$, a lower level of P210$^{c\text{-}abl}$ synthesis, or altered stability of the protein. To help distinguish among these possibilities, the steady-state level of P210$^{c\text{-}abl}$ in the cell lines was assayed by immunoblotting. The results show that SK-CML7Bt-33 (Fig. 2*A*, +) had a higher level of P210$^{c\text{-}abl}$ than P145, similar to the results with metabolic labeling (Fig. 1). We did not detect P210$^{c\text{-}abl}$ by immunoblotting with 2×10^7 cells of line SK-CML8Bt-3 (Fig. 2*B*, +). Reconstruction experiments using dilutions of cell extracts showed that we could detect about 5–10% the level of P210$^{c\text{-}abl}$ expressed in the K562 cell line (data not shown). We infer that the steady-state level of P210$^{c\text{-}abl}$ in SK-CML8Bt-3 is lower than the level in SK-CML7Bt-33 by a factor of at least 10. The level of P210$^{c\text{-}abl}$ detected in these assays correlated with the amount of P210$^{c\text{-}abl}$ tyrosine kinase activity that could be detected *in vitro* (data not shown).

Different Levels of P210$^{c\text{-}abl}$ Are Reflected in the Amount of Stable *bcr-abl* mRNA. To identify the basis for detection of variable levels of P210$^{c\text{-}abl}$, we examined the production of the *abl* RNA. RNA blot hybridization analysis using a v-*abl* probe (Fig. 3) showed that the normal 6- and 7-kb c-*abl* mRNAs were present at a similar level in Ph^1-positive and -negative cell lines derived from different patients. However, the 8-kb mRNA that encodes P210$^{c\text{-}abl}$ was detected at a 10-fold higher level in SK-CML7Bt-33 (Fig. 3*A*, +) than in SK-CML16Bt-1 (*B*, +), which correlated with the relative level of P210$^{c\text{-}abl}$ detected in each cell line. Analysis of additional cell lines demonstrated that the level of 8-kb RNA directly correlated with the level of P210$^{c\text{-}abl}$ (Table 1). The variation in level of 8-kb RNA detected in these cell lines was not due to loss or gain of Ph^1, because cytogenetic analysis confirmed the presence of Ph^1 in these cell lines (ref. 16 and

Table 1. Relative levels of *bcr-abl* expression in Epstein–Barr virus-immortalized B-cell lines and myeloid CML lines

Cell line*	CML phase†	Ph¹‡	P210§	8-kb mRNA¶
SK-CML7BN-2	BC	−	−	−
SK-CML8BN-10	Chronic	−	−	−
SK-CML8BN-12	Chronic	−	−	−
SK-CML16BN-1	Chronic	−	−	−
SK-CML35BN-1	Chronic	−	−	−
SK-CML7B5-33	BC	+	+++	+++
SK-CML21Bt-1	Acc	+	+++	+++
SK-CML21Bt-6	Acc	+	+++	+++
SK-CML8Bt-3	Chronic	+	+	±
SK-CML16Bt-1	Chronic	+	+	+
SK-CML35Bt-2	Chronic	+	+	+
K562	BC	+	+++++	+++++
BV173	BC	+	+++++	+++++
EM2	BC	+	+++++	+++++

*Cell lines derived from CML patients by transformation with Epstein–Barr virus as described (16). Names of cell lines indicate patient number and Ph¹ status: SK-CML7Bt indicates a cell line derived from patient 7 that carries the 9;22 Ph¹ translocation; N indicates a normal karyotype. Myeloid-erythroid cell lines (K562, EM2, and BV173) are described in previous publications (9, 11, 22, 33).

†Status of patient at the time cell line was derived. BC, blast crisis; Acc, accelerated phase.

‡Presence (+) or absence (−) of Ph¹ as demonstrated by karyotypic or Southern blot analysis.

§$P210^{c\text{-}abl}$ detected as described in legend to Fig. 1. B-cell lines derived from blast-crisis and accelerated-phase patients had levels of P210 3- to 5-fold higher (+++) than levels of P145. Chronic-phase-derived cell lines had P210 levels lower than or just equivalent (+) to the level of P145. Myeloid and erythroid lines had levels of P210 5- to 10-fold higher than P145 (+++++).

¶Eight-kilobase *bcr-abl* mRNA detected as described in legend to Fig. 2. Symbols: ±, borderline detectable; +++++, level of 8-kb mRNA 5- to 10-fold higher than that of the 6- and 7-kb c-*abl* mRNA species; +++, level of 8-kb mRNA 3- to 5-fold higher than that of the 6- and 7-kb species; +, a level approximately equivalent to that of the 6- and 7-kb messages.

data not shown). There was no difference in the copy number of *abl*-related sequences as judged by Southern blot analysis (Fig. 4). Only the K562 cell line control showed an amplification of *abl* sequences, as previously reported (22, 23). These combined data suggest that differential *bcr-abl* mRNA expression from a single gene template is responsible for the variable levels of $P210^{c\text{-}abl}$ detected. This could be mediated

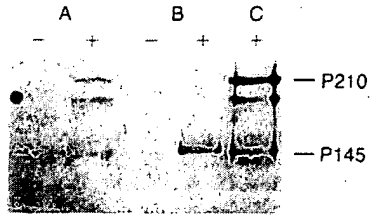


FIG. 2. Analysis of steady-state *abl* protein levels by immunoblotting. Cell extracts prepared from 2×10^7 cells of lines SK-CML7BN-2 (*A*,−), SK-CML7Bt-33 (*A*,+), SK-CML8BN-10 (*B*,−), and SK-CML8Bt-3 (*B*,+) were concentrated by immunoprecipitation with anti-pEX-2 plus anti-pEX-5. Samples were then electrophoresed in a $NaDodSO_4$/8% polyacrylamide gel and transferred to nitrocellulose, using protease-facilitated transfer (18). *abl* proteins were detected using a mixture of two monoclonal antibodies directed against the pEX-2 and pEX-5 *abl*-protein fragments produced in bacteria (19) as a probe and a peroxidase-conjugated goat anti-mouse second-stage antibody (Bio-Rad) for development.

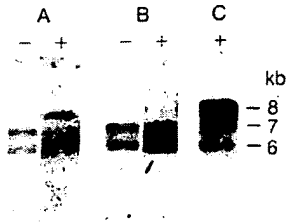


FIG. 3. Comparison of *abl* RNA levels in Ph¹-positive and -negative B-cell lines. The levels of the normal 6- and 7-kb c-*abl* RNAs and the 8-kb *bcr-abl* RNA were analyzed by blot hybridization using a v-*abl* probe. RNA was extracted from Ph¹-negative lines SK-CML7BN-2 (*A*,−) and SK-CML16BN-1 (*B*,−), from Ph¹-positive lines SK-CML6Bt-33 (*A*,+) and SK-CML16Bt-3 (*B*,+), and from line K562 (*C*,+) by the $NaDodSO_4$/urea/phenol method (20). Polyadenylylated RNA was purified by oligo(dT) affinity chromatography, and 15 μg of each sample was electrophoresed in a 1% agarose/formaldehyde gel and then transferred to nitrocellulose. The blotted RNAs were hybridized with a nick-translated v-*abl* fragment probe (21) and then autoradiographed for 4 days.

by factors influencing the transcription rate of the *bcr-abl* gene or the stability of the mRNA.

DISCUSSION

Several lines of evidence suggest that formation of Ph¹ is not the primary event that affects the stem cell in CML. Patients have been identified that present with the clinical picture of CML but only later develop Ph¹ (1). This observation, coupled with studies of *G6PD* (glucose-6-phosphate dehydrogenase)-heterozygous females with CML that demonstrate stem-cell clonality by isozyme analysis among cell

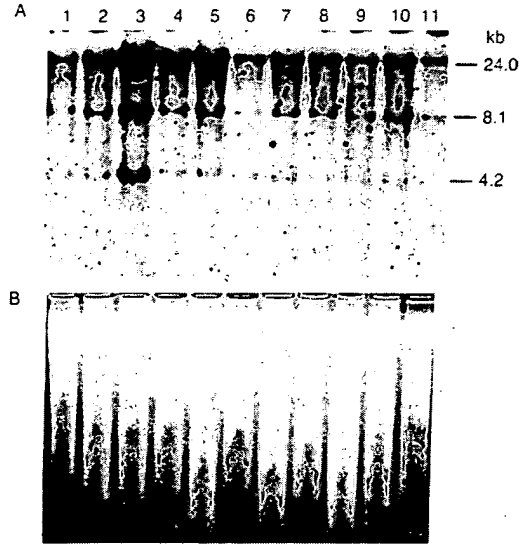


FIG. 4. Southern blot analysis of *abl* sequences in Ph¹-positive and -negative B-cell lines. High molecular weight DNA (15 μg) was digested with restriction endonuclease *Bam*HI, separated in a 0.8% agarose gel, and then transferred to nitrocellulose. The blotted DNA fragments were hybridized with a nick-translated, 2.4-kb *Bgl* II v-*abl* fragment (1.5×10^8 cpm/μg; ref. 21) and exposed for 4 days. (*A*) Autoradiogram of *abl*-specific fragments in cell lines HL-60 (lane 1), EM2 (lane 2), K562 (lane 3), SK-CML7Bt-33 (lane 4), SK-CML8Bt-3 (lane 5), SK-CML16Bt-1 (lane 6), SK-CML21Bt-6 (lane 7), SK-CML35Bt-2 (lane 8), SK-CML7BN-2 (lane 9), SK-CML8BN-2 (lane 10), and SK-CML35BN-1 (lane 11). (*B*) Ethidium bromide staining of agarose gel prior to transfer to nitrocellulose, showing the level of variation in amount of DNA loaded per lane.

populations that lack the Ph^1 marker, supports a secondary or complementary role for Ph^1 in the progression of the disease (24, 25). This chromosome marker is found in chronic, accelerated, and blast-crisis phases of the disease. It is likely that Ph^1 confers some growth advantage, since cells with the marker chromosome eventually predominate the marrow and peripheral blood even in chronic phase. During the phase of blast crisis, many patients develop additional chromosome abnormalities, including duplication of Ph^1, a variety of trisomies, and complex translocations (26). This is suggestive evidence for Ph^1 being a necessary but not sufficient genetic change for the full evolution of the disease.

The realization that one molecular result of Ph^1 is the generation of a chimeric *bcr-abl* protein with functional characteristics and structure analogous to the *gag-abl* transforming protein of the Abelson murine leukemia virus strengthens the argument for an important role of Ph^1 in the pathogenesis of CML. Although the Abelson virus is generally considered a rapidly transforming retrovirus, its effects can range from overcoming growth factor requirements, to cellular lethality, to induction of highly oncogenic tumors in a number of hematopoietic cell lineages (27, 28). Even in the transformation of murine cell targets, there are several lines of evidence that suggest that the growth-promoting activity of the v-*abl* gene product is complemented by further cellular changes in the production of the malignant-cell phenotype (29–31).

The regulation of *bcr-abl* gene expression is complex because the 5′ end of the gene is derived from the non-*abl* sequences, *bcr*, normally found on chromosome 22 (6). The level of stable message for the normal *bcr* gene and the normal *abl* gene are both much lower than the level of the *bcr-abl* message and protein from cell lines and clinical specimens derived from myeloid blast-crisis patients (5, 6, 11). Therefore, the high level of *bcr-abl* expression cannot simply be attributed to the regulatory sequences associated with *bcr*. Possibly, creation of the chimeric gene disrupts the normal regulatory sequences and results in a higher level of expression. Variation in *bcr-abl* expression may result from secondary changes in the structure of the chimeric gene or function of *trans*-acting factors that occur during evolution of the disease. Our analysis of $P210^{c\text{-}abl}$ and the 8-kb mRNA in Epstein–Barr virus-transformed Ph^1-positive B-cell lines demonstrates that stable message and protein levels from the *bcr-abl* gene can vary over a wide range. This variation does not result from a change in the number of *bcr-abl* templates secondary to gene amplification but more likely from changes in either transcription rate or mRNA stability. We suspect this range of *bcr-abl* expression is not limited to lymphoid cells. Analysis of peripheral blood leukocytes derived from an unusual CML patient who has been in chronic phase with myeloid predominance for 16 years showed a level of $P210^{c\text{-}abl}$ one-fifth that of $P145^{c\text{-}abl}$, as detected by metabolic labeling with [^{32}P]orthophosphate and immunoprecipitation (S.C., O.N.W., and P. Greenberg, unpublished observations). Lower levels of expression of the chimeric mRNA have been demonstrated in clinical samples from chronic-phase CML patients compared to acute-phase CML patients (9). Others have reported chronic-phase patients with variable but, in some cases, relatively high levels of the *bcr-abl* mRNA (32). The sampling variation and the heterogenous mixture of cell types in clinical samples complicate such analyses. Further work is needed to evaluate whether there is a defined change in $P210^{c\text{-}abl}$ expression during the progression of CML. It is interesting to note that among the limited sample of Ph^1-positive B-cell lines we have examined (Table 1), we have seen higher levels of $P210^{c\text{-}abl}$ in those derived from patients at more advanced stages of the disease. It will be important to search for cell-type-specific mechanisms that might regulate expression of *bcr-abl* from Ph^1.

We thank Bonnie Hechinger and Carol Crookshank for excellent secretarial assistance and Margaret Newman for excellent technical assistance. This work was supported by grants from the National Institutes of Health (to O.N.W. and B.C.). J.B.K. was supported as a predoctoral fellow on the Public Health Service Cellular and Molecular Biology Training Grant GM07185. S.C. is a postdoctoral fellow of the Leukemia Society of America.

1. Champlin, R. E. & Golde, D. W. (1985) *Blood* **65**, 1039–1047.
2. Rowley, J. D. (1973) *Nature (London)* **243**, 290–291.
3. Heisterkamp, N., Stephenson, J. R., Groffen, J., Hansen, P. F., de Klein, A., Bartram, C. R. & Grosveld, G. (1983) *Nature (London)* **306**, 239–242.
4. Bartram, C. R., de Klein, A., Hagemeijer, A., van Agthoven, T., van Kessel, A. G., Bootsma, D., Grosveld, G., Ferguson-Smith, M. A., Davies, T., Stone, M., Heisterkamp, N., Stephenson, J. R. & Groffen, J. (1983) *Nature (London)* **306**, 277–280.
5. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, D. (1985) *Nature (London)* **315**, 550–554.
6. Heisterkamp, N., Stam, K. & Groffen, J. (1985) *Nature (London)* **315**, 758–761.
7. Groffen, J., Stephenson, J. R., Heisterkamp, N., de Klein, A., Bartram, C. R. & Grosveld, G. (1984) *Cell* **36**, 93–99.
8. Gale, R. P. & Canaani, E. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5648–5652.
9. Collins, S., Kubonishi, L., Miyoshi, I. & Groudine, M. T. (1984) *Science* **225**, 72–74.
10. Konopka, J. B., Watanabe, S. M. & Witte, O. N. (1984) *Cell* **7**, 1035–1042.
11. Konopka, J. B., Watanabe, S. M., Singer, J., Collins, S. & Witte, O. N. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1810–1814.
12. Kloetzer, W., Kurzrock, R., Smith, L., Talpaz, M., Spiller, M., Gutterman, J. & Arlinghaus, R. (1985) *Virology* **140**, 230–238.
13. Kozbor, D., Giallongo, A., Sierzega, M. E., Konopka, J. B., Witte, O. N., Showe, L. C. & Croce, C. M. (1985) *Nature (London)*, in press.
14. Davis, R. L., Konopka, J. B. & Witte, O. N. (1985) *Mol. Cell Biol.* **5**, 204–213.
15. Konopka, J. B. & Witte, O. N. (1985) *Mol. Cell. Biol.* **5**, 3116–3123.
16. Nitta, M., Kato, Y., Strife, A., Wachter, M., Fried, J., Perez, A., Jhanwar, S., Duigou, R., Chaganti, R. S. K. & Clarkson, B. (1985) *Blood* **66**, 1053–1061.
17. Konopka, J. B., Davis, J. L., Watanabe, S. M., Ponticelli, A. S., Schiff-Maker, L., Rosenberg, N. & Witte, O. N. (1984) *Virology* **51**, 223–232.
18. Gibson, W. (1981) *Anal. Biochem.* **118**, 1–3.
19. Schiff-Maker, L., Konopka, J. B., Clark, S., Witte, O. N. & Rosenberg, N. (1986) *J. Virol.* **57**, 1182–1186.
20. Schwartz, R. C., Sonenshein, G. E., Bothwell, A. & Gefter, M. L. (1981) *J. Immunol.* **126**, 2104–2108.
21. Goff, S. P., Gilboa, E., Witte, O. N. & Baltimore, D. (1980) *Cell* **22**, 777–785.
22. Collins, S. J. & Groudine, M. T. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 4813–4817.
23. Selden, J. R., Emanuel, B. S., Wang, E., Cannizzaro, L., Palumbo, A., Erikson, J., Nowell, P. C., Rovera, G. & Croce, C. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7289–7292.
24. Fialkow, P. J., Martin, P. J., Najfeld, V., Penfold, G. K., Jacobson, R. J. & Hansen, J. A. (1981) *Blood* **58**, 158–163.
25. Martin, P. J., Najfeld, V. & Fialkow, P. J. (1982) *Can. Gen. Cytogenet.* **6**, 359–368.
26. Rowley, J. D. (1980) *Annu. Rev. Genet.* **14**, 17–40.
27. Whitlock, C. A. & Witte, O. N. (1984) *Adv. Immunol.* **37**, 74–98.
28. Pierce, J. H., Di Fiore, P. P., Aaronson, S. A., Potter, M., Pumphrey, J., Scott, A. & Ihle, J. N. (1985) *Cell* **41**, 685–693.
29. Whitlock, C. A., Ziegler, S. & Witte, O. N. (1983) *Mol. Cell. Biol.* **3**, 596–604.
30. Wolf, D., Harris, N. & Rotter, V. (1984) *Cell* **38**, 119–126.
31. Klein, G. & Klein, E. (1985) *Nature (London)* **315**, 190–195.
32. Stam, K., Jr., Heisterkamp, N., Grosveld, G., de Klein, A., Verma, R., Colemen, M., Dosik, H. & Groffen, J. (1985) *N. Engl. J. Med.* **313**, 1429–1433.
33. Lozzio, C. B. & Lozzio, B. B. (1975) *Blood* **45**, 321–334.

Review

Paul A. Haynes
Steven P. Gygi
Daniel Figeys
Ruedi Aebersold

Department of Molecular Biotechnology, University of Washington, Seattle, WA, USA

Proteome analysis: Biological assay or data archive?

In this review we examine the current state of proteome analysis. There are three main issues discussed: why it is necessary to study proteomes; how proteomes can be analyzed with current technology; and how proteome analysis can be used to enhance biological research. We conclude that proteome analysis is an essential tool in the understanding of regulated biological systems. Current technology, while still mostly limited to the more abundant proteins, enables the use of proteome analysis both to establish databases of proteins present, and to perform biological assays involving measurement of multiple variables. We believe that the utility of proteome analysis in future biological research will continue to be enhanced by further improvements in analytical technology.

Contents

1 Introduction

A proteome has been defined as the protein complement expressed by the genome of an organism, or, in multicellular organisms, as the protein complement expressed by a tissue or differentiated cell [1]. In the most common implementation of proteome analysis the proteins extracted from the cell or tissue analyzed are separated by high resolution two-dimensional gel electrophoresis (2-DE), detected in the gel and identified by their amino acid sequence. The ease, sensitivity and speed with which gel-separated proteins can be identified by the use of recently developed mass spectrometric techniques have dramatically increased the interest in proteome technology. One of the most attractive features of such analyses is that complex biological systems can potentially be studied in their entirety, rather than as a multitude of individual components. This makes it far easier to uncover the many complex, and often obscure, relationships between mature gene products in cells. Large-scale proteome characterization projects have been undertaken for a number of different organisms and cell types: Microbial proteome projects currently in progress include, for example: *Saccharomyces cerevisiae* [2], *Salmonella enterica* [3], *Spiroplasma melliferum* [4], *Mycobacterium tuberculosis* [5], *Ochrobactrum anthropi* [6], *Haemophilus influenzae* [7], *Synechocystis spp.* [8], *Escherichia coli* [9], *Rhizobium leguminosarum* [10], and *Dictyostelium discoideum* [11]. Proteome projects underway for tissues of more complex organisms include those for: human bladder squamous cell carcinomas [12], human liver [13], human plasma [13], human keratinocytes [12], human fibroblasts [12], mouse kidney [12], and rat serum [14]. In this manuscript we critically assess the concept of proteome analysis and the technical feasibility of establishing complete proteome maps, and discuss ways in which proteome analysis and biological research intersect.

Correspondence: Professor Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA, 98195, USA (Tel: +206-685-4235; Fax: +206-685-6392; E-mail: ruedi@u.washington.edu)

Abbreviations: CID, collision-induced dissociation; MS/MS, tandem mass spectrometry; SAGE, serial analysis of gene expression

Keywords: Proteome / Two-dimensional polyacrylamide gel electrophoresis / Tandem mass spectrometry

2 Rationale for proteome analysis

The dramatic growth in both the number of genome projects and the speed with which genome sequences are being determined has generated huge amounts of sequence information, for some species even complete genomic sequences ([15–17]). The description of the state of a biological system by the quantitative measurement of system components has long been a primary objective in molecular biology. With recent technical advances including the development of differential display-PCR [18], cDNA microarray and DNA chip technology [19, 20] and serial analysis of gene expression (SAGE) [21, 22], it is now feasible to establish global and quantitative mRNA expression maps of cells and tissues, in which the sequence of all the genes is known, at a speed and sensitivity which is not matched by current

0173-0835/98/1111-1862 $17.50+.50/0

protein analysis technology. Given the long-standing paradigm in biology that DNA synthesizes RNA which synthesizes protein, and the ability to rapidly establish global, quantitative mRNA expression maps, the questions which arise are why technically complex proteome projects should be undertaken and what specific types of information could be expected from proteome projects which cannot be obtained from genomic and transcript profiling projects. We see three main reasons for proteome analysis to become an essential component in the comprehensive analysis of biological systems. (i) Protein expression levels are not predictable from the mRNA expression levels, (ii) proteins are dynamically modified and processed in ways which are not necessarily apparent from the gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

2.1 Correlation between mRNA and protein expression levels

Interpretations of quantitative mRNA expression profiles frequently implicitly or explicitly assume that for specific genes the transcript levels are indicative of the levels of protein expression. As part of an ongoing study in our laboratory, we have determined the correlation of expression at the mRNA and protein levels for a population of selected genes in the yeast *Saccharomyces cerevisiae* growing at mid-log phase (S. P. Gygi *et al.*, submitted for publication). mRNA expression levels were calculated from published SAGE frequency tables [22]. Protein expression levels were quantified by metabolic radiolabeling of the yeast proteins, liquid scintillation counting of the protein spots separated by high resolution 2-DE and mass spectrometric identification of the protein(s) migrating to each spot. The selected 80 samples constitute a relatively homogeneous group with respect to predicted half-life and expression level of the protein products. Thus far, we have found a general trend but no strong correlation between protein and transcript levels (Fig. 1). For some genes studied equivalent mRNA transcript levels translated into protein abundances which varied by more than 50-fold. Similarly, equivalent steady-state protein expression levels were maintained by transcript levels varying by as much as 40-fold (S. P. Gygi *et al.*, submitted). These results suggests that even for a population of genes predicted to be relatively homogeneous with respect to protein half-life and gene expression, the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript.

2.2 Proteins are dynamically modified and processed

In the mature, biologically active form many proteins are post-translationally modified by glycosylation, phosphorylation, prenylation, acylation, ubiquitination or one or more of many other modifications [23] and many proteins are only functional if specifically associated or complexed with other molecules, including DNA, RNA, proteins and organic and inorganic cofactors. Frequently, modifications are dynamic and reversible and may alter the precise three-dimensional structure and the state of activity of a protein. Collectively, the state of modification of the proteins which constitute a biological system

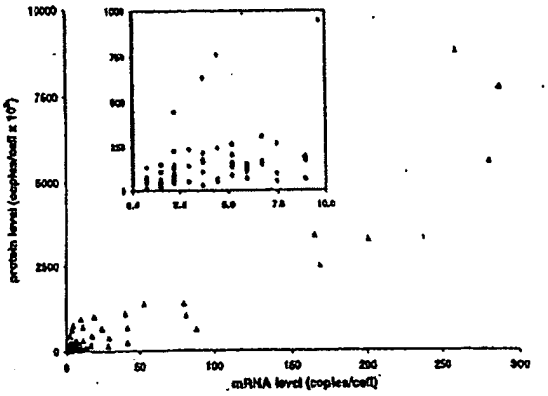


Figure 1. Correlation between mRNA and protein levels in yeast cells. For a selected population of 80 genes, protein levels were measured by ^{35}S-radiolabeling and mRNA levels were calculated from published SAGE tables. Inset: expanded view of the low abundance region. For more experimental details, also see Figs. 5 and 6, (S. P. Gygi *et al.*, submitted).

are important indicators for the state of the system. The type of protein modification and the sites modified at a specific cellular state can usually not be determined from the gene sequence alone.

2.3 Proteomes are dynamic and reflect the state of a biological system

A single genome can give rise to many qualitatively and quantitatively different proteomes. Specific stages of the cell cycle and states of differentiation, responses to growth and nutrient conditions, temperature and stress, and pathological conditions represent cellular states which are characterized by significantly different proteomes. The proteome, in principle, also reflects events that are under translational and post-translational control. It is therefore expected that proteomics will be able to provide the most precise and detailed molecular description of the state of a cell or tissue, provided that the external conditions defining the state are carefully determined. In answer to the question of whether the study of proteomes is necessary for the analysis of biomolecular systems, it is evident that the analysis of mature protein products in cells is essential as there are numerous levels of control of protein synthesis, degradation, processing and modification, which are only apparent by direct protein analysis.

3 Description and assessment of current proteome analysis technology

3.1 Technical requirements of proteome technology

In biological systems the level of expression as well as the states of modification, processing and macro-molecular association of proteins are controlled and modulated depending on the state of the system. Comprehensive analysis of the identity, quantity and state of modification of proteins therefore requires the detection and

quantitation of the proteins which constitute the system, and analysis of differentially processed forms. There are a number of inherent difficulties in protein analysis which complicate these tasks. First, proteins cannot be amplified. It is possible to produce large amounts of a particular protein by over-expression in specific cell systems. However, since many proteins are dynamically post-translationally modified, they cannot be easily amplified in the form in which they finally function in the biological system. It is frequently difficult to purify from the native source sufficient amounts of a protein for analysis. From a technological point of view this translates into the need for high sensitivity analytical techniques. Second, many proteins are modified and processed post-translationally. Therefore, in addition to the protein identity, the structural basis for differentially modified isoforms also needs to be determined. The distribution of a constant amount of protein over several differentially modified isoforms further reduces the amount of each species available for analysis. The complexity and dynamics of post-translational protein editing thus significantly complicates proteome studies. Third, proteins vary dramatically with respect to their solubility in commonly used solvents. There are few, if any, solvent conditions in which all proteins are soluble and which are also compatible with protein analysis. This makes the development of protein purification methods particularly difficult since both protein purification and solubility have to be achieved under the same conditions. Detergents, in particular sodium dodecyl sulfate (SDS), are frequently added to aqueous solvents to maintain protein solubility. The compatibility with SDS is a big advantage of SDS polyacrylamide gel electrophoresis (SDS-PAGE) over other protein separation techniques. Thus, SDS-PAGE and two-dimensional gel electrophoresis, which also uses SDS and other detergents, are the most general and preferred methods for the purification of small amounts of proteins, provided that activity does not necessarily need to be maintained. Lastly, the number of proteins in a given cell system is typically in the thousands. Any attempt to identify and categorize all of these must use methods which are as rapid as possible to allow completion of the project within a reasonable time frame. Therefore, a successful, general proteomics technology requires high sensitivity, high throughput, the ability to differentiate differentially modified proteins, and the ability to quantitatively display and analyze all the proteins present in a sample.

3.2 2-D electrophoresis – mass spectrometry: a common implementation of proteome analysis

The most common currently used implementation of proteome analysis technology is based on the separation of proteins by two-dimensional (IEF/SDS-PAGE) gel electrophoresis and their subsequent identification and analysis by mass spectrometry (MS) or tandem mass spectrometry (MS/MS). In 2-DE, proteins are first separated by isoelectric focusing (IEF) and then by SDS-PAGE, in the second, perpendicular dimension. Separated proteins are visualized at high sensitivity by staining or autoradiography, producing two-dimensional arrays of proteins. 2-DE gels are, at present, the most commonly used means of global display of proteins in complex samples. The separation of thousands of proteins has been achieved in a single gel [24, 25] and differentially modified proteins are frequently separated. Due to the compatibility of 2-DE with high concentrations of detergents, protein denaturants and other additives promoting protein solubility, the technique is widely used.

The second step of this type of proteome analysis is the identification and analysis of separated proteins. Individual proteins from polyacrylamide gels have traditionally been identified using *N*-terminal sequencing [26, 27], internal peptide sequencing [28, 29], immunoblotting or comigration with known proteins [30]. The recent dramatic growth of large-scale genomic and expressed sequence tag (EST) sequence databases has resulted in a fundamental change in the way proteins are identified by their amino acid sequence. Rather than by the traditional methods described above, protein sequences are now frequently determined by correlating mass spectral or tandem mass spectral data of peptides derived from proteins, with the information contained in sequence databases [31–33].

There are a number of alternative approaches to proteome analysis currently under development. There is considerable interest in developing a proteome analysis stragegy which bypasses 2-DE altogether, because it is considered a relatively slow and tedious process, and because of perceived difficulties in extracting proteins from the gel matrix for analysis. However, 2-DE as a starting point for proteome analysis has many advantages compared to other techniques available today. The most significant strengths of the 2-DE-MS approach include the relatively uniform behavior of proteins in gels, the ability to quantify spots and the high resolution and simultaneous display of hundreds to thousands of proteins within a reasonable time frame.

A schematic diagram of a typical procedure of the identification of gel-separated proteins is shown in Fig. 2. Protein spots detected in the gel are enzymatically or chemically fragmented and the peptide fragments are isolated for analysis, as already indicated, most frequently by MS or MS/MS. There are numerous protocols for the generation of peptide fragments from gel-separated proteins. They can be grouped into two categories, digestion in the gel slice [28, 34] or digestion after electrotransfer out of the gel onto a suitable membrane ([29, 35–37] and reviewed in [38]). In most instances either technique is applicable and yields good results. The analysis of MS or MS/MS data is an important step in the whole process because MS instruments can generate an enormous amount of information which cannot easily be managed manually. Recently, a number of groups have developed software systems dedicated to the use of peptide MS and MS/MS spectra for the identification of proteins. Proteins are identified by correlating the information contained in the MS spectra of protein digests or MS/MS spectra of individual peptides with data contained in DNA or protein sequence databases.

The systems we are currently using in our laboratory are based on the separation of the peptides contained in protein digests by narrow bore or capillary liquid chromatog-

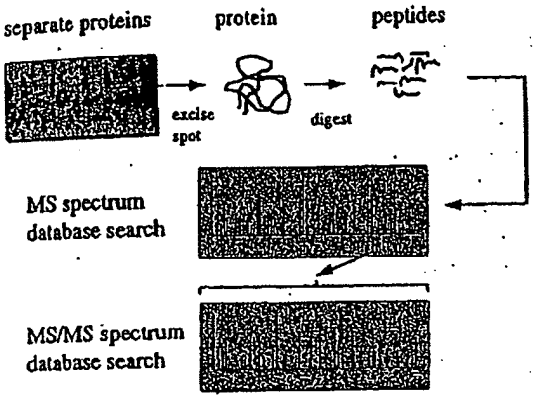


Figure 2. Schematic diagram of a procedure for identification of gel-separated proteins. Peptides can either be separated by a technique such as LC or CE, or infused as a mixture and sorted in the MS. Database searching can either be performed on peptide masses from an MS spectrum, peptide fragment masses from CID spectra of peptides, or a combination of both.

raphy [39, 40] or capillary electrophoresis [41], the analysis of the separated peptides by electrospray ionization (ESI) MS/MS, and the correlation of the generated peptide spectra with sequence databases using the SEQUEST program developed at the University of Washington [32, 33]. The system automatically performs the following operations: a particular peptide ion characterized by its mass-to-charge ratio is selected in the MS out of all the peptide ions present in the system at a particular time; the selected peptide ion is collided in a collision cell with argon (collision-induced dissociation, CID) and the masses of the resulting fragment ions are determined in the second sector of the tandem MS; this experimentally determined CID spectrum is then correlated with the CID spectra predicted from all the peptides in a sequence database which have essentially the same mass as the peptide selected for CID; this correlation matches the isolated peptide with a sequence segment in a database and thus identifies the protein from which the peptide was derived. There are a number of alternative programs which use peptide CID spectra for protein identification, but we use the SEQUEST system because it is currently the most highly automated program and has proven to be successful, versatile and robust.

3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS

It has been demonstrated repeatedly that MS has a very high intrinsic sensitivity. For the routine analysis of gel-separated proteins at high sensitivity, the most significant challenge is the handling of small amounts of sample. The crux of the problem is the extraction and transferal of peptide mixtures generated by the digestion of low nanogram amounts of protein, from gels into the MS/MS system without significant loss of sample or introduction of unwanted contaminants. We employ three different systems for introducing gel-purified samples into an MS, depending on the level of sensitivity required. As an approximate guideline, for samples containing tens of picomoles of peptides, LC-MS/MS is most appropriate; for samples containing low picomole amounts to high femtomole amounts we use capillary LC-MS/MS; and for samples containing femtomoles or less, CE-MS/MS is the method of choice.

3.3.1 LC-MS/MS

The coupling of an MS to an HPLC system using a 0.5 mm diameter or bigger reverse phase (RP) column has been described in detail [42]. This system has several advantages if a large number of samples are to be analyzed and all are available in sufficient quantity. The LC-MS and database searching program can be run in a fully automated mode using an autosampler, thus maximizing sample throughput and minimizing the need for operator interference. The relatively large column is tolerant of high levels of impurities from either gel preparation or sample matrix. Lastly, if configured with a flow-splitter and micro-sprayer [40], analyses can be performed on a small fraction of the sample (less than 5%) while the remainder of the sample is recovered in very pure solvents. This latter feature is particularly useful when an orthogonal technique is also used to analyze peptide fractions, such as scintillation of an introduced radiolabel, and this data can be correlated with peptides identified by CID spectra.

3.3.2 Capillary LC-MS

An increase of sensitivity of approximately tenfold can be achieved by using a capillary LC system with a 100 μm ID column rather than a 0.5 mm ID column as referred to above. Since very low flow rates are required for such columns, most reports have used a precolumn flow splitting system for producing solvent gradients. We have recently desribed the design and construction of a novel gradient mixing system which enables the formation of reproducible gradients at very low flow rates (low nL/min) without the need for flow splitting (A. Ducret *et al.*, submitted for publication). Using this capillary LC-MS/MS system we were able to identify gel-separated proteins if low picomole to high femtomole amounts were loaded onto the gel [40]. This system is as yet not automated and, like all capillary LC systems, is prone to blockage of the columns by microparticulates when analyzing gel-separated proteins.

3.3.3 CE-MS/MS

The highest level of sensitivity for analyzing gel-separated proteins can be achieved by using capillary electrophoresis – mass spectrometry (CE-MS). We have described in the past a solid-phase extraction capillary electrophoresis (SPE-CE) system which was used with triple quadrupole and ion trap BSI-MS/MS systems for the identification of proteins at the low femtomole to sub-femtomole sensitivity level [43, 44]. While this system is highly sensitive, its operation is labor-intensive and its operation has not been automated. In order to devise an analytical system with both the sensitivity of a CE and the level of automation of LC, we have constructed

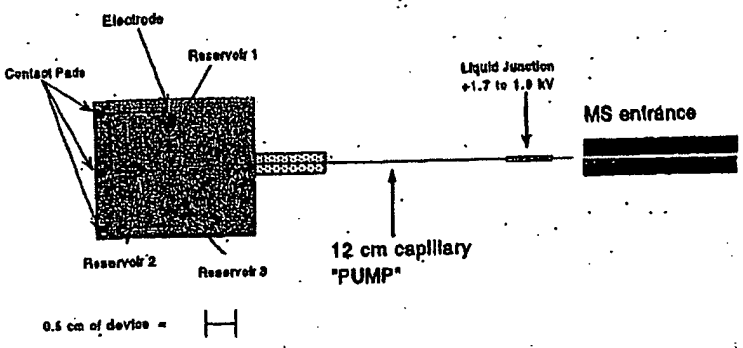


Figure 3. Schematic illustration of a microfabricated analytical system for CE, consisting of a micromachined device, coated capillary electroosmotic pump, and microelectrospray interface. The dimensions of the channels and reservoir are as indicated in the text. The channels on the device were graphically enhanced to make them more visible. Reproduced from [45], with permission.

microfabricated devices for the introduction of samples into ESI-MS for high-sensitivity peptide analysis.

The basic device is a piece of glass into which channels of 10–30 μm in depth and 50–70 μm in diameter are etched by using photolithography/etching techniques similar to the ones used in the semiconductor industry. (A simple device is shown in Fig. 3). The channels are connected to an external high voltage power supply [45]. Samples are manipulated on the device and off the device to the MS by applying different potentials to the reservoirs. This creates a solvent flow by electroosmotic pumping which can be redirected by changing the position of the electrode. Therefore, without the need for valves or gates and without any external pumping, the flow can be redirected by simply switching the position of the electrodes on the device. The direction and rate of the flow can be modulated by the size and the polarity of the electric field applied and also by the charge state of the surface.

The type of data generated by the system is illustrated in Fig. 4, which shows the mass spectrum of a peptide sample representing the tryptic digest of carbonic anhydrase at 290 fmol/μL. Each numbered peak indicates a peptide successfully identified as being derived from carbonic anhydrase. Some of the unassigned signals may be chemical or peptide contaminants. The MS is programmed to automatically select each peak and subject the peptide to CID. The resulting CID spectra are then used to identify the protein by correlation with sequence databases. Therefore, this system allows us to concurrently apply a number of protein digests onto the device, to sequentially mobilize the samples, to automatically generate CID spectra of selected peptide ions and to search sequence databases for protein identification. These steps are performed automatically without the need for user input and proteins can be identified at very low femtomole level sensitivity at a rate of approximately one protein per 15 min.

3.4 Assessment of 2-DE-MS proteome technology

Using a combination of the analytical techniques described above we have identified the 80 protein spots indicated in Fig. 5. The protein pattern was generated by separating a total of 40 microgram of protein contained in a total cell lysate of the yeast strain YPH499 by high resolution 2-DE and silver staining of the separated proteins. To estimate how far this type of proteome analysis can penetrate towards the identification of low abundance proteins, we have calculated the codon bias of the genes encoding the respective proteins. Codon bias is a

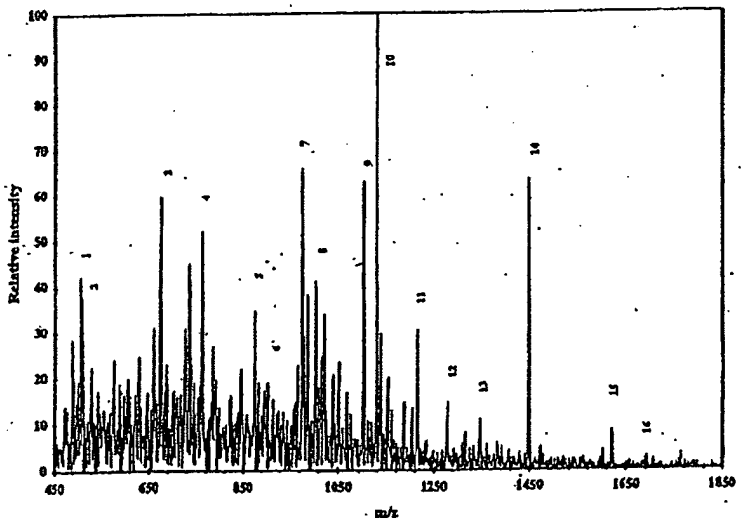


Figure 4. MS spectrum of a tryptic digest of carbonic anhydrase using the microfabricated system shown in Fig. 3. 290 fmol/μL of carbonic anhydrase tryptic digest was infused into a Finnigan LCQ ion trap MS. Each peak was selected for CID, and those which were identified as containing peptides derived from carbonic anhydrase are numbered. Reproduced from [45], with permission.

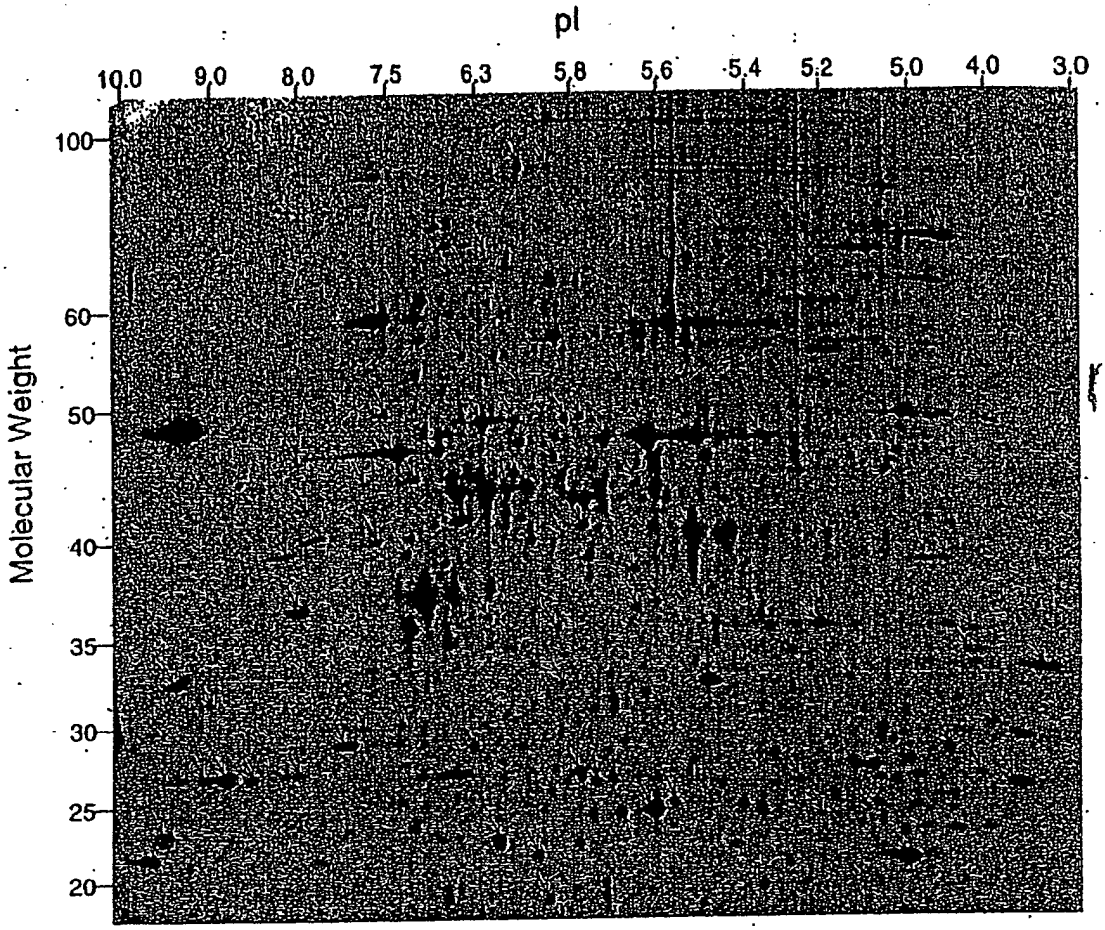


Figure 5. 2-DE separation of a lysate of yeast cells, with identified proteins highlighted. The first dimension of separation was an IPG from pH 3–10, and the second dimension was a 10%T SDS-PAGE gel. Proteins were visualized by silver staining. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

calculated measure of the degree of redundancy of triplet DNA codons used to produce each amino acid in a particular gene sequence. It has been shown to be a useful indicator of the level of the protein product of a particular gene sequence present in a cell [46]. The general rule which applies is that the higher the value of the codon bias calculated for a gene, the more abundant the protein product of that gene becomes. The calculated codon bias values corresponding to the proteins identified in Fig. 5 are shown in Fig. 6b. Nearly all of the proteins identified (> 95%) have codon bias values of > 0.2, indicating they are highly abundant in cells. In contrast, codon bias values calculated for the entire yeast genome (Fig. 6a) show that the majority of proteins present in the proteome have a codon bias of < 0.2 and are thus of low abundance.

This finding is of considerable importance in our assessment of the current status of proteome analysis technology. It is clear that even using highly sensitive analytical techniques, we are only able to visualize and identify the more abundant proteins. Since many important regulatory proteins are present only at low abundance, these would not be amenable to analysis using such techniques. This situation would be exacerbated in the analysis of proteomes containing many more proteins than the approximately 6000 gene products present in yeast cells [16]. In the analysis of, for example, the proteome of any human cells; there are potentially 50 000–100 000 gene products [47]. Inherent limitations on the amount of protein that can be loaded on 2-DE, and the number of components that can be resolved, indicate that only the most highly abundant fraction of the many gene products could be successfully analyzed. One approach that has been employed to circumvent these limitations is the use of very narrow range immobilized pH gradient strips for the first-dimension separation of 2-DE [48]. Since only those proteins which focus within the narrow range will enter the second dimension of separation, a much higher sample loading within the desired range is possible. This, in turn, can lead to the visualization and identification of less abundant proteins.

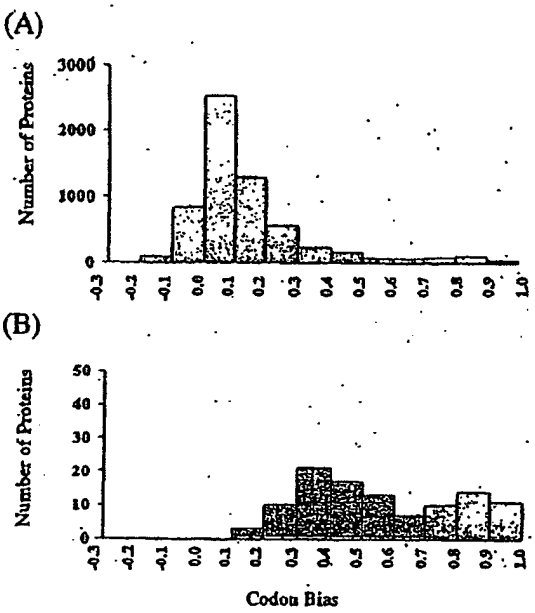


Figure 6. Calculated codon bias values for yeast proteins. (A) Distribution of calculated values for the entire yeast proteome. (B) Distribution of calculated values for the subset of 80 identified proteins also shown in Figs. 1 and 5. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

4 Utility of proteome analysis for biological research

For the success of proteomics as a mainstream approach to the analysis of biological systems it is essential to define how proteome analysis and biological research projects intersect. Without a clear plan for the implementation of proteome-type approaches into biological research projects the full impact of the technology can not be realized. The literature indicates that proteome analysis is used both as a database/data archive, and as a biological assay or biological research tool.

4.1 The proteome as a database

The use of proteomics as a database or data archive essentially entails an attempt to identify all the proteins in a cell or species and to annotate each protein with the known biological information that is relevant for each protein. The level of annotation can, of course, be extensive. The most common implementation of this idea is the separation of proteins by high resolution 2-DE, the identification of each detected protein spot and the annotation of the protein spots in a 2-DE gel database format. This approach is complicated by the fact that it is difficult to precisely define a proteome and to decide which proteome should be represented in the database. In contrast to the genome of a species, which is essentially static, the proteome is highly dynamic. Processes such as differentiation, cell activation and disease can all significantly change the proteome of a species. This is illustrated in Fig. 7. The figure shows two high-resolution 2-DE maps of proteins isolated from rat serum. Fig. 7A is from the serum of normal rats, while Fig. 7B is from the serum of rats in acute-phase serum after prior treatment with an inflammation-causing agent [49]. It is obvious that the protein patterns are significantly different in several areas, raising the question of exactly which proteome is being described.

Therefore, a comprehensive proteome database of a species or cell type needs to contain all of the parameters which describe the state and the type of the cells from which the proteins were extracted as well as the software tools to search the database with queries which reflect the dynamics of biological systems. A comprehensive proteome database should be capable of quantitatively describing the fate of each protein if specific systems and pathways are activated in the cell. Specifically, the quantity, the degree of modification, the subcellular location and the nature of molecules specifically interacting with a protein as well as the rate of change of these variables should be described. Using these admittedly stringent criteria, there is currently no comlete proteome database. A number of such databases are, however, in the process of being constructed. The most advanced among them, in our opinion, are the yeast protein database YPD [50] (accessible at http://www.ypd.com) and the human 2D-PAGE databases of the Danish Centre for Human Genome Research [12] (accessible at http://biobase.dk/cgi-bin/celis). While neither can be considered complete as not all of the potential gene products are identified, both contain extensive annotation of supplemental information for many of the spots which are positively identified in reference samples.

4.2 The proteome as a biological assay

The use of proteome analysis as a biological assay or research tool represents an alternative approach to integrating biology with proteomics. To investigate the state of a system, samples are subjected to a specific proceess that allows the quantitative or qualitative measurement of some of the variables which describe the system. In typical biochemical assays one variable (*e.g.*, enzyme activity) of a single component (*e.g.*, a particular enzyme) is measured. Using proteomics as an assay, multiple variables (*e.g.*, expression level, rate of synthesis, phosphorylation state, etc.) are measured concurrently on many (ideally all) of the proteins in a sample. The use of proteomics as an assay is a less far-reaching proposition than the construction of a comprehensive proteome database. It does, however, represent a pragmatic approach which can be adapted to investigate specific systems and pathways, as long as the interpretation of the results takes into account that with current technology not all of the variables which describe the system can be observed (see Section 3.4).

A common implementation of proteome analysis as a biological assay is when a 2-DE protein pattern generated from the analysis of an experimental sample is compared to an array of reference patterns representing different states of the system under investigation. The state of the experimental system at the time the sample was generated is therefore determined by the quantita-

tive comparative analysis of hundreds to a few thousand proteins. Comparative analysis of the 2-DE patterns furthermore highlights quantitative and qualitative differences in the protein profiles which correlate with the state of the system. For this type of analysis it is not essential that all the proteins are identified or even visualized, although the results become more informative as more proteins are compared. It is obvious, however, that the possibility to identify any protein deemed characteristic for a particular state dramatically enhances this approach by opening up new avenues for experimentation.

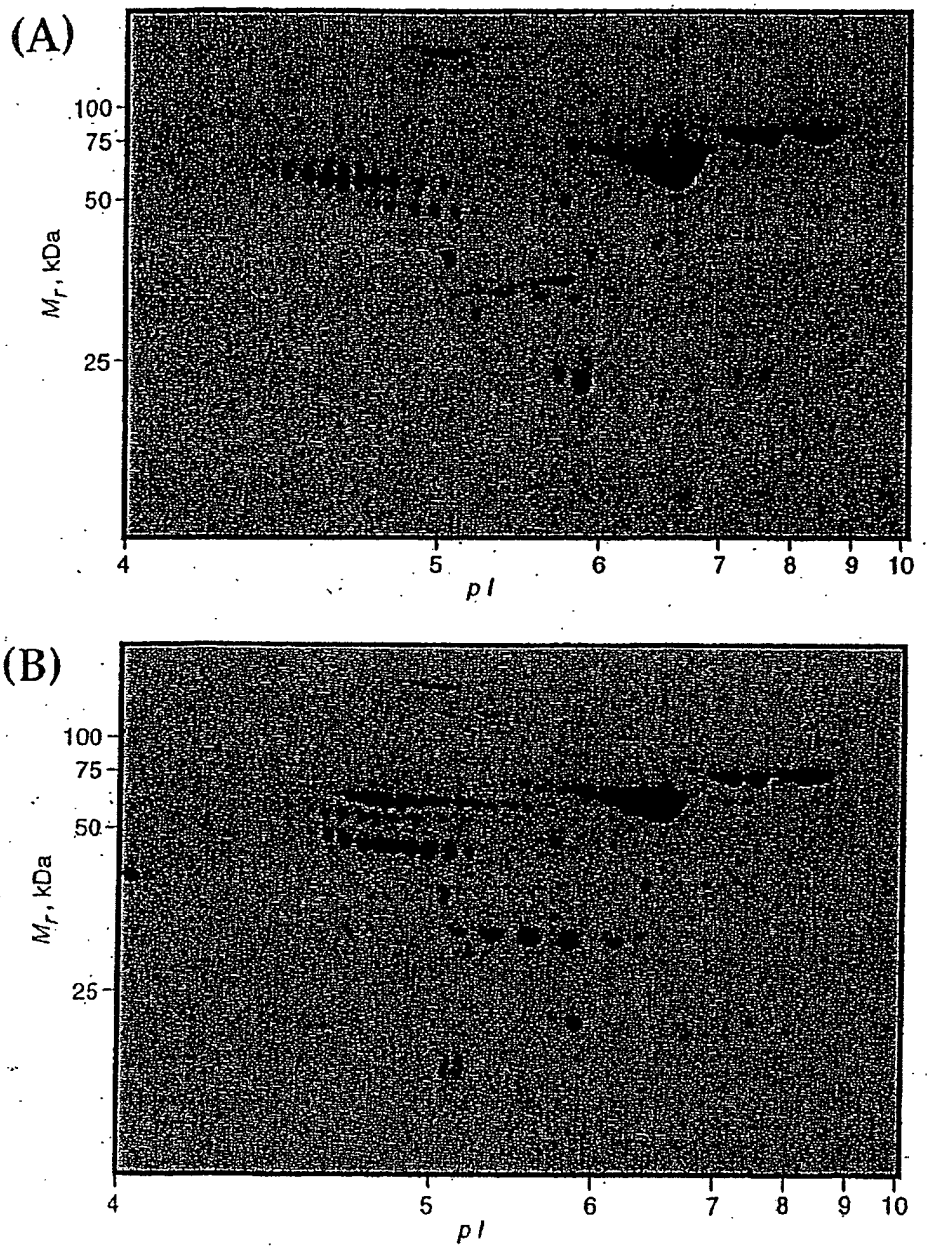


Figure 1. High resolution 2-DE map of proteins isolated from rat serum with or without prior exposure to an inflammation-causing agent. (A) normal rat serum, (B) acute-phase serum from rats which had previously been exposed to an inflammation-causing agent. The first dimension of separation is an IPG from pH 4–10, and the second dimension is a 7.5–17.5%T gradient SDS-PAGE gel. Proteins were visualized by staining with amido black. Further details of experimental procedures are included in [14, 49].

Proteome analysis as a biological assay has been successfully used in the field of toxicology, to characterize disease states or to study differential activation of cells. The approach is limited, of course, by the fact that only the visible protein spots are included in the assay, and it is well known that a substantial but far from complete fraction of cellular proteins are detected if a total cell lysate is separated by 2-DE. Proteins may not be detected in 2-DE gels because they are not abundant enough to be visualized by the detection method used, because they do not migrate within the boundaries (size, *pI*) resolved by the gel, because they are not soluble under the conditions used, or for other reasons.

A different way to use proteome analysis as a biological assay to define the state of a biological system is to take advantage of the wealth of information contained in 2-DE protein patterns. 2-DE is referred to as two-dimensional because of the electrophoretic mobility and the isoelectric points which define the position of each protein in a 2-DE pattern. In addition to the two dimensions used to generate the protein patterns, a number of additional data dimensions are contained in the protein patterns. Some of these dimensions such as protein expression level, phosphorylation state, subcellular location, association with other proteins, rate of synthesis or degradation indicate the activity state of a protein or a biological system. Comparative analysis of 2-DE protein patterns representing different states is therefore ideally suited for the detection, identification and analysis of suitable markers. Once again it must be emphasized that in this type of experiment only a fraction of the cellular proteins is analyzed. Since many regulatory proteins are of low abundance, this limitation is a concern, particularly in cases in which regulatory pathways are being investigated.

5 Concluding remarks

In this report we have addressed three main issues related to proteome analysis. First, we have discussed the rationale for studying proteomes. Second, we have assessed the technical feasibility of analyzing proteomes and described current proteome technology, and third, we have analyzed the utility of proteome analysis for biological research. It is apparent that proteome analysis is an essential tool in the analysis of biological systems. The multi-level control of protein synthesis and degradation in cells means that only the direct analysis of mature protein products can reveal their correct identities, their relevant state of modification and/or association and their amounts. Recently developed methods have enabled the identification of proteins at ever-increasing sensitivity levels and at a high level of automation of the analytical processes. A number of technical challenges, however, remain. While it is currently possible to identify essentially any protein spots that can be visualized by common staining methods, it is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions.

We have outlined the two principal ways proteome analysis is currently being used to intersect with biological research projects: the proteome as a database or data archive and proteome analysis as a biological assay. Both approaches have in common that at present they are conceptually and technically limited. Current proteome databases typically are limited to one cell type and one state of a cell and therefore do not account for the dynamics of biological systems. The use of proteome analysis as a biological assay can provide a wealth of information, but it is limited to the proteins detected and is therefore not truly proteome-wide. These limitations in proteomics are to a large extent a reflection of the fact that proteins in their fully processed form cannot easily be amplified and are therefore difficult to isolate in amounts sufficient for analysis or experimentation. The fact that to date no complete proteome has been described further attests to these difficulties. With continued rapid progress in protein analysis technology, however, we anticipate that the goal of complete proteome analysis will eventually become attainable.

We would like to acknowledge the funding for our work from the National Science Foundation Science and Technology Center for Molecular Biotechnology and from the NIH. We thank Yvan Rochon and Bob Franza for providing the yeast gel shown and Elisabetta Gianazza for providing the rat serum gels shown.

Received April 21, 1998

6 References

[1] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., *Bio/Technology* 1996, *14*, 61–65.

[2] Hodges, P. E., Payne, W. E., Garrels, J. I., *Nucleic Acids Res.* 1998, *26*, 68–72.

[3] O'Connor, C. D., Ferris, M., Fowler, R., Qi, S. Y., *Electrophoresis* 1997, *18*, 1483–1490.

[4] Cordwell, S. J., Basseal, D. J., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1335–1346.

[5] Urquhart, B. L., Atzafos, T. E., Roach, D., Basseal, D. J., Bjellqvist, B., Britton, W. L., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1384–1392.

[6] Wasinger, V. C., Bjellqvist, B., Humphery-Smith, I., *Electrophoresis* 1997, *18*, 1373–1383.

[7] Link, A. J., Hays, L. G., Carmack, E. B., Yates III, J. R., *Electrophoresis* 1997, *18*, 1314–1334.

[8] Sazuka, T., Ohara, O., *Electrophoresis* 1997, *18*, 1252–1258.

[9] VanBogelen, R. A., Abshire, K. Z., Moldover, B., Olson, E. R., Neidhardt, F. C., *Electrophoresis* 1997, *18*, 1243–1251.

[10] Guerreiro, N., Redmond, J. W., Rolfe, B. G., Djordjevic, M. A., *Mol. Plant Microbe Interact.* 1997, *10*, 506–516.

[11] Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Gooley, A. A., Hochstrasser, D. F., Williams, K. L., *Electrophoresis* 1997, *18*, 491–497.

[12] Celis, J., Gromov, P., Ostergaard M., Madsen, P., Honoré, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., Rasmussen, H. H., *FEBS Lett.* 1996, *398*, 129–134.

[13] Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, *14*, 1232–1238.

[14] Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, R. M., Manzoni, C., Vignati, M., Gianazza, E., *Electrophoresis* 1998, *19*, 1484–1492.

[15] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., Venter, J. C., *Science* 1995, *269*, 496–512.

[16] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., *Science* 1996, *274*, 546.

[17] Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, T., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., Venter, J. C., *Nature* 1997, *390*, 580–586.

[18] Liang, P., Pardee, A. B., *Science* 1992, *257*, 967–971.

[19] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, *94*, 13057–13062.

[20] Shalon, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, *6*, 639–645.

[21] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, *270*, 484–487.

[22] Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., Kinzler, K. W., *Cell* 1997, *88*, 243–251.

[23] Krishna, R. G., Wold, F., *Adv. Enzymol.* 1993, *67*, 265–298.

[24] Görg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, *9*, 531–546.

[25] Klose, J., Kobalz, U., *Electrophoresis* 1995, *16*, 1034–1059.

[26] Matsudaira, P., *J. Biol. Chem.* 1987, *262*, 10035–10038.

[27] Aebersold, R. H., Teplow, D. B., Hood, L. E., Kent, S. B., *J. Biol. Chem.* 1986, *261*, 4229–4238.

[28] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, *203*, 173–179.

[29] Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B., *Proc. Natl. Acad. Sci. USA* 1987, *84*, 6970–6974.

[30] Honoré, B., Leffers, H., Madsen, P., Celis, J. B., *Eur. J. Biochem.* 1993, *218*, 421–430.

[31] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[32] Eng, J., McCormack, A. L., Yates III, J. R., *J. Amer. Mass Spectrom.* 1994, *5*, 976–989.

[33] Yates III, J. R., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.

[34] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[35] Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., Aebersold, R., *Protein Sci.* 1993, *2*, 1342–1351.

[36] van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Biochem.* 1997, *247*, 310–318.

[37] Lui, M., Tempst, P., Brdjument-Bromage, H., *Anal. Biochem.* 1996, *241*, 156–166.

[38] Patterson, S. D., Aebersold, R. A., *Electrophoresis* 1995, *16*, 1791–1814.

[39] Ducret, A., Foyn, Brunn, C., Bures, E. J., Marhaug, G., Husby, G. R. A., *Electrophoresis* 1996, *17*, 866–876.

[40] Haynes, P. A., Fripp, N., Aebersold, R., *Electrophoresis* 1998, *19*, 939–945.

[41] Figeys, D., Van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, *68*, 1822–1828.

[42] Ducret, A., Van Oostveen, I., Eng, J. K., Yates III, J. R., Aebersold, R., *Protein Sci.* 1997, *7*, 706–719.

[43] Figeys, D., Ducret, A., Yates III, J. R., Aebersold, R., *Nature Biotech.* 1996, *14*, 1579–1583.

[44] Figeys, D., Aebersold, R., *Electrophoresis* 1997, *18*, 360–368.

[45] Figeys, D., Ning, Y., Aebersold, R., *Anal. Chem.* 1997, *69*, 3153–3160.

[46] Garrels, J. I., McLaughlin, C. S., Warner, J. R., Futcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., Payne, W. E., *Electrophoresis* 1997, *18*, 1347–1360.

[47] Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, X., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voytlicky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehejia, A., Goodfellow, P. N., Houlgatte, R., Hudson, J. R., Jr., Ide, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., *et al. Science* 1996, *274*, 540–546.

[48] Sanchez, J.-C., Rouge, V., Pisteur, M., Ravier, F., Tonella, L., Moosmayer, M., Wilkins, M. R., Hochstrasser, D. F., *Electrophoresis* 1997, *18*, 324–327.

[49] Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., Gianazza, E., *Electrophoresis* 1998, *19*, 1493–1500.

[50] Garrels, J. I., *Nucleic Acids Res.* 1996, *24*, 46–49.

research articles Journal of proteome research

Analysis of Genomic and Proteomic Data Using Advanced Literature Mining

Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera, Andrea Richardson, and Joshua LaBaer*

Institute of Proteomics, Harvard Medical School–BCMP, 240 Longwood Avenue, Boston, Massachusetts 02115

Received March 13, 2003

High-throughput technologies, such as proteomic screening and DNA micro-arrays, produce vast amounts of data requiring comprehensive analytical methods to decipher the biologically relevant results. One approach would be to manually search the biomedical literature; however, this would be an arduous task. We developed an automated literature-mining tool, termed MedGene, which comprehensively summarizes and estimates the relative strengths of all human gene–disease relationships in Medline. Using MedGene, we analyzed a novel micro-array expression dataset comparing breast cancer and normal breast tissue in the context of existing knowledge. We found no correlation between the strength of the literature association and the magnitude of the difference in expression level when considering changes as high as 5-fold; however, a significant correlation was observed ($r = 0.41$; $p = 0.05$) among genes showing an expression difference of 10-fold or more. Interestingly, this only held true for estrogen receptor (ER) positive tumors, not ER negative. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative tumors worthy of further examination.

Keywords: bioinformatics • micro-array • text mining • gene-disease association • breast cancer

Introduction

At its current pace, the accumulation of biomedical literature outpaces the ability of most researchers and clinicians to stay abreast of their own immediate fields, let alone cover a broader range of topics. For example, to follow a single disease, e.g., breast cancer, a researcher would have had to scan 130 different journals and read 27 papers per day in 1999.[1] This problem is accentuated with high-throughput technologies such as DNA micro-arrays and proteomics, which require the analysis of large datasets involving thousands of genes, many of which are unfamiliar to a particular researcher. In any microarray experiment, thousands of genes may demonstrate statistically significant expression changes, but only a fraction of these may be relevant to the study. The ability to interpret these datasets would be enhanced if they could be compared to a comprehensive summary of what is known about all genes. Thus, there is a need to summarize existing knowledge in a format that allows for the rapid analysis of associations between genes and diseases or other specific biological concepts.

One solution to this problem is to compile structured digital resources, such as the Breast Cancer Gene Database[1] and the Tumor Gene Database.[2] However, as these resources are hand-curated, the labor-intensive review process becomes a rate-limiting step in the growth of the database. As a result, these databases have a limited scale and the genes are not selected in a systematic fashion.

An alternative approach is automated text mining; a method which involves automated information extraction by searching documents for text strings and analyzing their frequency and context. This approach has been used successfully in several instances for biological applications. In most cases, it has been applied to extract information about the relationships or interactions that proteins or genes have with one another, in the literature or by functional annotation.[3–7] Thus far, few publication have applied text-mining to examine the global relationships between genes and diseases. Perez-Iratxeta et al. automatically examined the GO (Gene Ontology) annotation of genes and their predicted chromosomal locations in order to identify genes linked to inherited disorders.[8]

To obtain a more global understanding of disease development, it would be valuable to incorporate information regarding all possible gene-disease relationships, including biochemical, physiological, pharmacological, epidemiological, as well as genetic. This information would enable comprehensive comparisons between large experimental datasets and existing knowledge in the literature. This would accomplish two things. First, it would serve to validate experiments by demonstrating that known responses occur as predicted. Second, it would rapidly highlight which genes are corroborated by the literature and which genes are novel in a given context. We have utilized a computational approach to literature mining to produce a

* To whom correspondence should be addressed: jlabaer@hms.harvard.edu.

10.1021/pr0340227 CCC: $25.00

Published on Web 06/13/2003

comprehensive set of gene-disease relationships. In addition, we have developed a novel approach to assess the strength of each association based on the frequency of citation and co-citation. We applied this tool to help interpret the data from a large micro-array gene expression experiment comparing normal and cancerous breast tissue.

Methods

MedGene Database. MedGene is a relational database, storing disease and gene information from NCBI, text mining results, statistical scores, and hyperlinks to the primary literature. MedGene has a web-based user interface for users to query the database (http://hipseq.med.harvard.edu/MedGene/).

Text Mining Algorithms. MeSH files were downloaded from the MeSH web site at NLM (Nation Library of Medicine) (http://www.nlm.nih.gov/mesh/meshhome.html) and human disease categories were selected. LocusLink files were downloaded from the LocusLink web site at NCBI (http://www.ncbi.nih.gov/LocusLink/). Official/preferred gene symbol, official/preferred gene name, and gene alternative symbols and names, all relevant annotations and URLs for each LocusLink record, were collected. Gene search terms were used for literature searching and included all qualified gene names, gene symbols, and gene family terms. Primary gene keys, predominantly qualified gene family terms and gene official/preferred symbols, were used to index Medline records. If the official/preferred gene symbols did not meet the standards to be an index, then qualified gene official/preferred names were used. A local copy of Medline records (up to July, 2002) was pre-selected.

A JAVA module examined the MeSH terms and then indexed each Medline record with the appropriate disease terms. A separate JAVA module was used to examine the titles and abstracts for gene search terms and then to index the gene-related Medline records with the relevant primary gene key(s).

Statistical Methods. For every gene and disease pair, we counted records that were indexed for both gene and disease (double positive hits), for disease only (disease single hits), for gene only (gene single hits), and for neither gene nor disease (double negative hits) to generate a 2 × 2 contingency table. On the basis of the contingency table-framework, we applied different statistical methods to estimate the strength of gene-disease relationships and evaluated the results. These methods included chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease[16] (http://hipseq.med.harvard.edu/MedGene/). In addition, we computed the "product of frequency", which is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. To obtain a normal distribution, we transformed all the statistical scores using the natural logarithm. We selected the log of the product of frequency (LPF) to validate MedGene and to use for the analysis with the micro-array data. Spearman rank-correlation coefficients were used to assess the linear relationship between LPF and micro-array fold change in expression level.

Global Analysis. Diseases with at least 50 related genes were selected for clustering analysis, and the LPF scores were normalized with total score for each disease. Hierarchical clustering was done with the "Cluster" software and the clustering result was visualized using "TreeViewer" (http://rana.lbl.gov/EisenSoftware.htm).

Breast Tissue Micro-Arrays. Eighty-nine breast cancer samples (79% ER-positive) and 7 normal breast tissue samples were selected from the Harvard Breast SPORE frozen tissue repository and were representative of the spectrum of histological types, grades, and hormone receptor immuno-phenotypes of breast cancer. Biotinylated cRNA, generated from the total RNA extracted from the bulk tumor, was hybridized to Affymetrix U95A oligo-nucleotide micro-arrays. These micro-arrays consist of 12 400 probes, which represent approximately 9000 genes. Raw expression values were obtained using GENECHIP software from Affymetrix, and then further analyzed using the DNA-Chip Analyzer (dChip) custom software.

Results

Automated Indexing of Medline Records by Disease and Gene. To study the gene-disease associations in the literature, we first compiled complete lists for human diseases and human genes. To index all Medline records that were relevant to human diseases, the Medical Subject Heading (MeSH) index of Medline records was utilized. MeSH is a controlled medical vocabulary from the National Library of Medicine and consists of a set of terms or subject headings that are arranged in both an alphabetic and an hierarchical structure. Medline records are reviewed manually and MeSH terms are added to each with software assistance.[9,10] Twenty-three human disease category headings along with all of their child terms (see the Supporting Information, Supplemental Table 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 1.html) were selected from the 2002 MeSH index creating a list of 4033 human diseases.

No index comparable to the MeSH index exists for genes, and thus, it was necessary to apply a string search algorithm for gene names or symbols found in Medline text. A complete list of genes, gene names, gene symbols, and frequently used synonyms were collected from the LocusLink database at NCBI,[11,12] which contains 53 259 independent records keyed by an official gene symbol or name (June 18th, 2002). For the purposes of this study, no distinction was made between genes and their gene products. Authors often use the same name for both, differentiating the two only by the use of italics, if at all. For the intended use of this study, this lack of distinction is unlikely to have a large effect and may in fact be beneficial.

Initial attempts to search the literature using these lists revealed several sources of false positives and false negatives (Table 1). False positives primarily arose when the searched term had other meanings, whereas false negatives arose from syntax discrepancies necessitating the development of filters to reduce these errors. The syntax issues were readily handled by including alternate syntax forms in the search terms. The false positive cases, caused by duplicative and unrelated meanings for the terms, were more difficult to manage. Where possible, case sensitive string mapping reduced inappropriate citations. In many cases, however, this was not sufficient and the terms had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.

For the purposes of data tracking, a primary gene key was selected to represent all synonyms that correspond to each gene. Medline records were indexed with a primary gene key when any synonym for that key was found in the title or abstract. Case-insensitive string mapping was used for all searches except as noted above. No additional weight was

Table 1. Systematic Sources of False Positives and False Negatives in Unfiltered Data[a]

source of error	error type	example	filter solution
gene symbol/name is not unique	false positive	*MAG*—myelin associated glycoprotein *MAG*—malignancy-associated protein	eliminate this term
gene symbol is unrelated abbreviation	false positive	*PA*—pallid homologue (mouse), pallidin (also abbrev. for Pennsylvania)	eliminate this term
gene symbol/name has language meaning	false positive	*WAS*—Wiskott–Aldrich Syndrome (also the word "was")	case-sensitive string search
nonstandard syntax	false negative	*BAG-1* instead of *BAG1*	add dash term
unofficial gene name/symbol	false negative	*P53* instead of *TP53*	add all gene nicknames
nonspecified gene name	false negative	estrogen receptor instead of Estrogen receptor 1	add family stem term

[a] In preliminary studies, Medline was searched for co-occurrence of genes and diseases and the resulting output was evaluated to identify error sources that were amenable to global filters. Each error source is categorized by the type of error it causes: false positives are suggested relationships that are not real and false negatives are real relationships that are underrepresented. The filter solutions used are indicated. Note that in some cases, the filter solution itself introduces error. In general, error rates maximized sensitivity, even at the expense of specificity if needed.

added for multiple occurrences of a term or the co-occurrence of multiple synonyms for the same gene key.

Medline records were searched with all qualified gene identifiers, such as the official/preferred gene symbol, the official/preferred gene name, all gene nicknames and all syntax variants. In situations where there are several members of a gene family or splice variants, some authors prefer to use a shortened gene family name, e.g., estrogen receptor instead of estrogen receptor 1 (*ESR1*), creating a source of false negatives. For this reason, gene family stem terms were created for all genes that have an alpha or numerical suffix (e.g., *IL2RA*, *TGFβ*, *ESR1*, etc.) and then used to search the literature. The family stem terms were handled separately from the specific gene names so that it would be clear when linkages were made to the gene family versus a specific member in that family.

To improve performance and accuracy, some pre-selection was applied to the records that were scanned. First, review articles were eliminated to avoid redundant treatment of citations. Second, non-English journals were removed because the natural language filters were only relevant to English publications. Finally, journals unlikely to contain primary data about gene-disease relationships were also removed (e.g., *Int. J. Health Educ.*, *Bedside Nurse*, and *J. Health Econ.*). Together, these filters reduced the 12 198 221 Medline publications (July 2002) by 37%.

Ranking the Relative Strengths of Gene-Disease Associations. In total, there were 618 708 gene-disease co-citations, in which 16% (8297) of all studied genes had been associated to a disease and 96% (3875) of all diseases had been associated to at least one gene. To rank the relative strengths of gene disease relationships, we tested several different statistical methods and examined the results. With the exception of the relative risk estimates, the methods provided similar results with respect to the rank order of the gene-disease association strengths. However, after comparing the results to other databases and after consulting disease experts, the log of the product of frequency (LPF) was selected for further analysis because it gave the best results overall.

Validation of MedGene. In developing this tool, it was important to minimize the number of missed genes (false negatives) and miscalled genes (false positives). However, in situations when these goals were in conflict, inclusiveness was prioritized. To determine the false negative rate in MedGene, breast cancer was used as a test case because it was associated with more genes than any other human disease and because

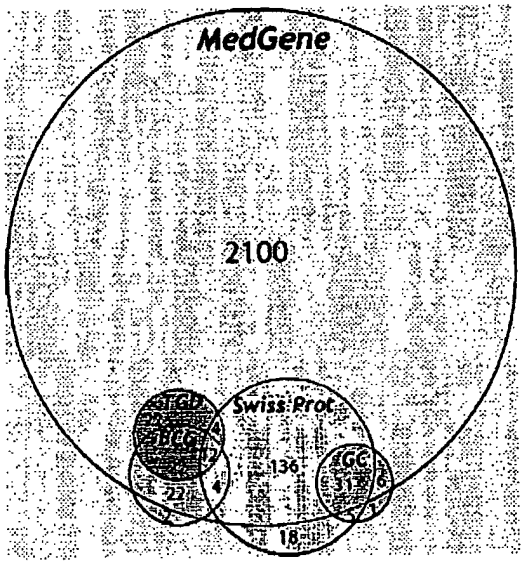


Figure 1. Estimation of the false negative rate by comparison with hand-curated databases. The breast cancer-related genes identified by MedGene were compared with those listed in several other databases including the Tumor Gene Database (TGD),[2] the Breast Cancer Gene Database(BCG),[1] GeneCards (GC)[17] and Swissprot.[18] Genes were considered false negatives if they were represented in at least one of these other databases and not in MedGene and their link to breast cancer was supported by at least one literature reference. All literature references were verified by manual review to confirm their validity. The number of genes in each database or shared by more than one database is indicated. The false negative rate was calculated by genes missed at MedGene (26)/total number of nonoverlapping genes in other databases (285).

there were several public databases that link genes to breast cancer. We compared the list of breast cancer-related genes from MedGene to these databases, illustrated in Figure 1. Among the 285 distinct breast cancer-related genes that were supported by at least one literature citation in these hand-curated databases, 26 were absent from MedGene, suggesting a false negative rate of approximately 9%. To determine why these were missed, all literature references for these genes (80

papers) were reviewed manually (see the Supporting Information, Supplemental Table 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 2.html). Among these papers, most false negatives were caused by nonstandard gene terms or gene terms eliminated by our specificity filters. Few genes were missed because they were only mentioned in review papers (0.4%) or they appeared only in the body of the manuscript but not the abstract or title (1.1%). Of note, MedGene identified approximately 2000 additional breast cancer-related genes not listed in any other database.

To assess the false positive error rate, two complementary approaches were used: a detailed analysis of one disease and a global examination of 1000 diseases. The detailed approach examined the false positive error rate and its sources, whereas the global approach tested whether the overall results made biomedical sense.

Using the LPF, 1467 genes related to prostate cancer were assembled in rank order. We then retrieved approximately 300 Medline records each for the highest ranked 100 and the lowest ranked 200 genes and manually reviewed the titles and abstracts to determine the verity of the association. Nearly 80% of the highest ranked 100 genes fell into one of the five categories that reflect meaningful gene-disease relationships (see the Supporting Information, Supplemental Table 3, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Among the lowest ranked 200 genes, approximately 70% reflected true relationships. Of the 600 records reviewed, there were only two in which the association between the gene and the disease was described as negative. Both were genes with very low scores. In both cases, the authors did not argue the absence of *any* relationship, but rather that a particular feature of the gene or protein was not shown to be related to human prostate cancer.[13,14]

The coincidence of some gene symbols with medical abbreviations, chemical abbreviations and biological abbreviations resulted in most of the false positives (see the Supporting Information, Supplemental Table 4, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Table 4.html), emphasizing the importance of the filters that were added in the search algorithm (Table 1). Without the filters, the false positive rate more than doubled, and the false negative rate rose dramatically (data not shown). For example, among the papers about breast cancer, there were only 12 Medline records that referred to *ESR1* and 10 to *ESR2*, whereas almost 2000 papers mentioned estrogen receptor without specifying *ESR1* or *ESR2*; this latter group was detected by the family stem term filter.

To further validate these results, a global analysis of the gene-disease relationships described by MedGene was performed. For this experiment, it was reasoned that the more closely related the diseases are to one another, the more they will be related to the same gene sets. Thus, if the relationships defined by MedGene accurately reflected the literature, then an unsupervised hierarchical clustering of the gene data should group diseases in a manner consistent with common medical thinking. Conversely, if the clustered diseases do not make sense biologically or medically, it may reflect excessive false positives, false negatives, or inappropriate scoring of the data.

To execute this experiment, the gene sets and the corresponding LPF values for 1000 randomly selected diseases (each with at least 50 gene relationships) were used as a dataset for clustering the diseases. A review of the results showed that the resulting disease clusters were indeed logical based upon common medical knowledge (see the Supporting Information, Supplemental Figure 1, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html). For example, in one such cluster shown in Figure 2, diabetes and its complications grouped together and were also closely linked to diseases associated with starvation states.

The number of genes associated with a given disease can be estimated by adjusting the MedGene number up by the false negative rate (~9%) and down by the false positive rate (~26% on average). Using this, the average disease has 103.7 ± 45.3 (mean ± s.d.) genes associated with it, although the range is quite broad with 2359 genes related to breast cancer, 2122 genes related to lung cancer and no genes related to a number of diseases.

Applying MedGene to the Analysis of Large Datasets. Access to a comprehensive summary of the genes linked to human diseases provided an opportunity to analyze data obtained from a high-throughput experiment. We compared the MedGene breast cancer gene list to a gene expression data set generated from a micro-array analysis comparing breast cancer and normal breast tissue samples. Micro-array analysis identified 2286 genes that had greater than a 1-fold difference in mean expression level between breast cancer samples and normal breast samples. Using MedGene, we sorted the 2286 genes into four classes: 555 genes directly linked to breast cancer in the literature by gene term search (first-degree association by gene name); 328 genes directly linked by family term search (first-degree association by family term); 1021 genes linked to breast cancer only through other breast cancer genes (second-degree association); and 505 genes not previously associated with breast cancer. (See the Supporting Information, Supplemental Figure 2, or visit http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 2.html.) Among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease. Among the remaining 38 genes, 9 had been related to other cancers, specifically esophageal, colon, uterine, skin, and cervix.

To determine whether the genes highlighted by the micro-array analysis were more likely to have been previously linked to breast cancer in the literature, we created a two-dimensional plot of the fold change of expression level between breast cancer and normal tissue versus the literature score (LPF) (Figure 3A). There was a broad spread of expression changes among the genes directly linked to breast cancer ranging from less than 1-fold change (68%) to over 40-fold (0.3%). Notably, the majority of genes with greater than 10-fold expression changes were linked to breast cancer by first-degree association.

Among all 754 genes directly linked to breast cancer in the literature, there was no correlation between LPF and micro-array fold change ($r = 0.018$, p-value = 0.62). However, when we stratified the analysis based on the magnitude of the fold change, we observed an increasing trend in correlation (Figure 3B) suggesting that genes with a more substantial change in expression level were more likely to have a stronger association in the literature. For genes that had 10-fold change or more in expression level, the correlation increased to 0.41 (p-value = 0.05).

When we evaluated the micro-array data separately for ER positive and ER negative tumors, the trend in correlation between fold change and literature score was highly dependent on estrogen receptor status. Interestingly, there was a similar trend in correlation for ER positive tumors, but no trend in correlation for ER negative tumors.

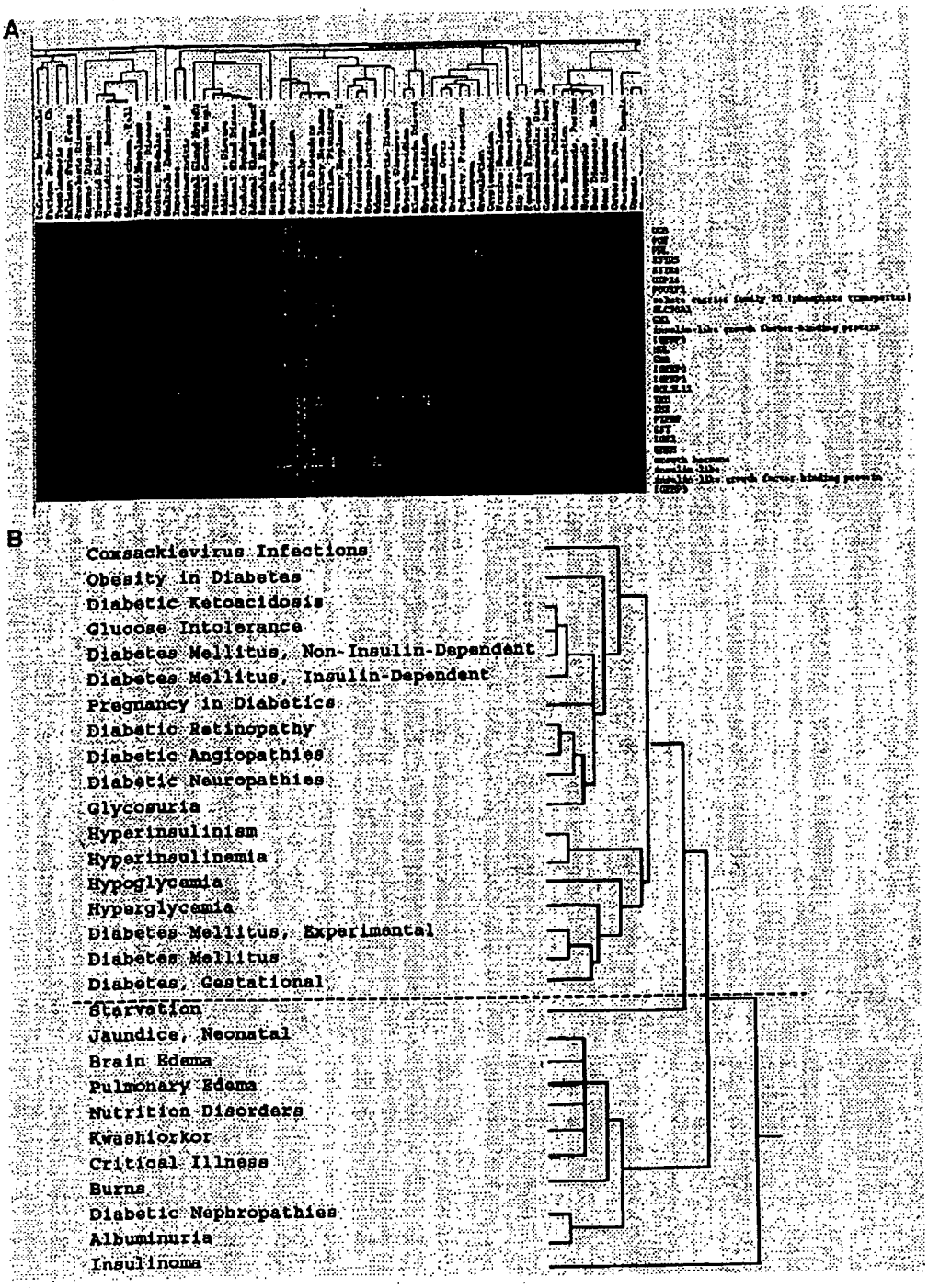


Figure 2. Global validation by clustering analysis. 2(A). The gene sets and the corresponding LPF values for 1000 diseases, each with at least 50 gene relationships, were used in an unsupervised clustering of the diseases based on the gene patterns associated with them. A sample of the data is shown here. 2(B). One of the resulting clusters is shown that corresponds to blood sugar states. Diabetes terms (above the line) and starvation states terms (under the line) clustered together. Within these groups, there is also clustering of diabetic small vessel complications, altered serum chemistries, nutritional disorders, etc.(Supplemental Figure 1: http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html).

Finally, to validate our findings, we computed similar correlations between the breast cancer expression data and LPF scores generated by MedGene for hypertension, a disease unrelated to breast cancer. As expected, we did not observe an increasing trend in correlation for hypertension.

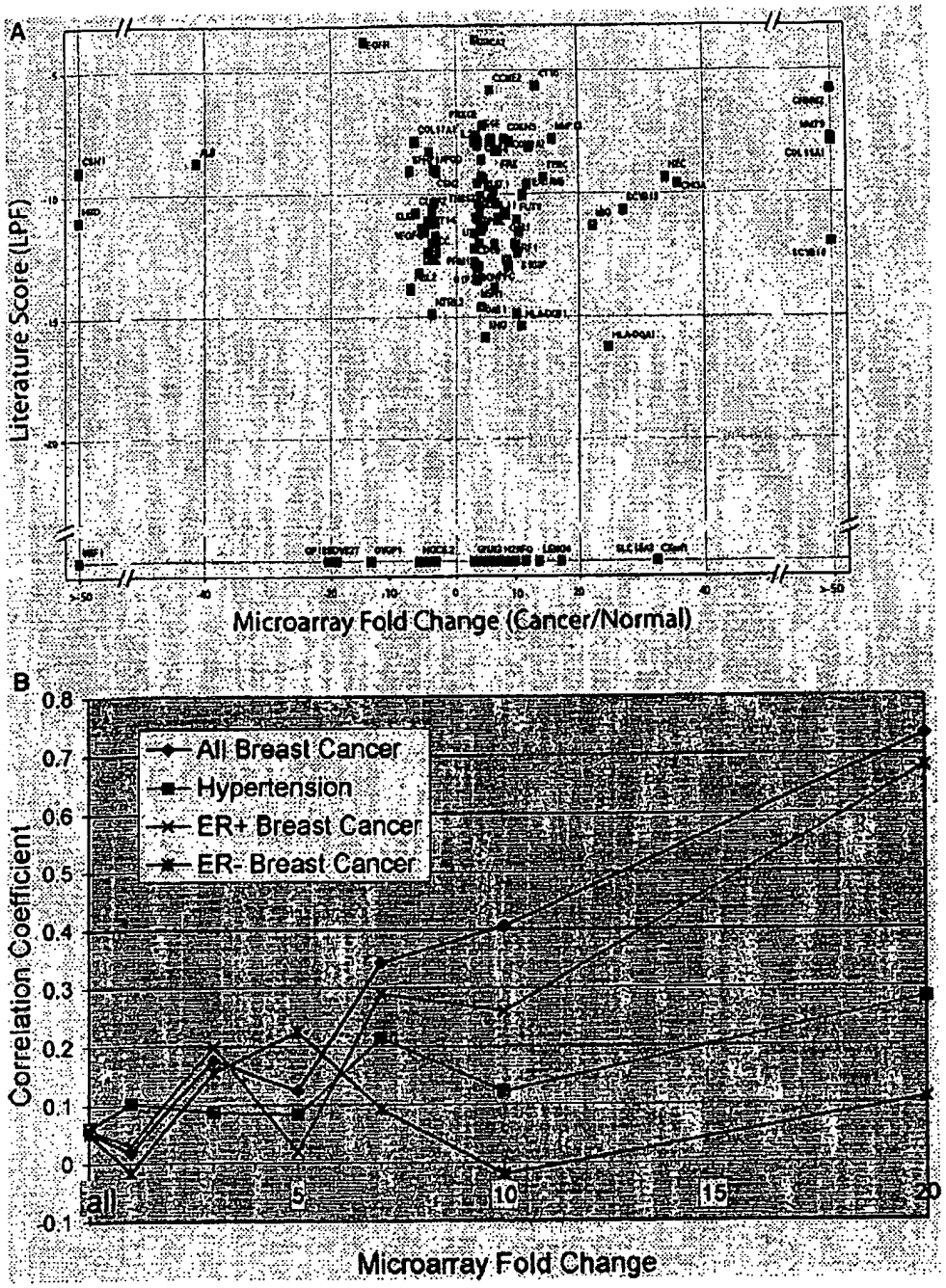


Figure 3. Relationship between literature score and functional data for breast cancer. 3A. The data from an expression analysis of samples for breast tumors and normal breast tissue were analyzed to indicate the fold difference of expression level between breast tumor and normal sample (cutoff ≥ 3-fold change). The fold changes were plotted against the literature score for the same gene set. Green dots represent first-degree association by gene search, blue dots represent first-degree association by family search and red dots represent no-association. Some well-studied genes, such as BRCA2 (pink circle), are not reflected by a substantial difference in expression level. Furthermore, the majority of genes that have no association with breast cancer in the literature had less than 10-fold expression changes (shaded area). 3B. The Spearman rank-correlation coefficients between literature score (LPF) and the fold change of expression level between tumor and normal breast samples (*y*-axis) in relation to the amount of fold change of expression level (*x*-axis). Gene rank lists were generated for breast cancer (blue) and hypertension (pink). Correlations were also computed between the breast cancer gene LPF scores and fold change expression data among estrogen receptor positive tumors only (light blue) and estrogen receptor negative tumors only (purple).

Table 2. Top 25 Genes Related to Selected Human Diseases[a]

breast neoplasms	hypertension	rheumatoid arthritis	bipolar disorder	atherosclerosis
estrogen receptor	*REN*	*RA*	*ERDA1*	apolipoprotein
PGR	*DBP*	*TNFRSF10A*	*SNAP29*	*APOE*
ERBB2	*LEP*	*CRP*	*PFKL*	*LDLR*
BRCA1	*AGT*	*AS*	*DRD2*	*ELN*
BRCA2	*INS*	*ESR1*	*TRH*	*ARG1*
EGFR	kallikrein	*HLA-DRB1*	*IMPA2*	*APOB*
CYP19	*ACE*	*DR1*	*HTR3A*	*APOA1*
TFF1	endothelin	interleukin	*DRD3*	*MSR1*
PSEN2	*S100A6*	*TNF*	*REM*	*LPL*
TP53	*BDK*	*IL6*	*KCNN3*	*PON1*
				plasminogen
CES3	*DIANPH*	collagen	*DRD4*	activator inhibitor
CEACAM5	*SAR1*	*IL1A*	*HTR2C*	*PLG*
				vascular cell
ERBB3	*PIH*	*ACR*	*RELN*	adhesion molecule
cyclin	*CD59*	*TNFRSF12*	*DBH*	*ATOH1*
COX5A	*ALB*	*IL2*	*MAOA*	*VWF*
cathepsin	*CYP11B2*	*CHI3L1*	*COMT*	*INS*
ERBB4	*MAT2B*	*IL8*	*HTR2A*	*ARG2*
	angiotensin			
TRAM	receptor	interleukin 1	*SYNJ1*	*ABCA1*
		matrix		
CCND1	*AGTR2*	metalloproteinase	*INPP1*	*OLR1*
EGF	*NPPA*	interferon	*NEDD4L*	collagen
MUC1	*LVM*	*CD68*	*FRA13C*	*MCP*
			transducer of	
insulin-like	*DBH*	*IL4*	*ERBB2*	lipoprotein
BCL2	*NPY*	*IL17*	*BAIAP3*	*APOA2*
				intercellular
mucin	*POMC*	*MMP3*	*ATP1B3*	adhesion molecule
FGF3	neuropeptide	*SIL*	*DRD5*	*RAB27A*

[a] MedGene results for the top 25 genes associated with breast neoplasms, hypertension, rheumatoid arthritis, bipolar disorder, and atherosclerosis, respectively, ranked by LPF scores. The hyperlink to all the papers co-citing the gene and the disease is available at MedGene website (http://hipseq.med.harvard.edu/MedGene/).

Discussion

The Human Genome Project heralded a new era in biological research where the emphasis on understanding specific pathways has expanded to global studies of genomic organization and biological systems. High-throughput technologies can provide novel insight into comprehensive biological function but also introduces new challenges. The utility of these technologies is limited to the ability to generate, analyze, and interpret large gene lists. MedGene, a relational database derived by mining the information in Medline, was created to address this need. MedGene users can query for a rank-ordered list of human gene-disease relationships (Table 2) for one or more diseases. Each entry is hyperlinked to the original papers supporting each association and to other relevant databases.

MedGene is an innovative extension of previous text mining approaches. Perez-Iratxeta et al. used the GO annotation and their chromosomal locations to predict genes that may contribute to inherited disorders.[6] MedGene takes a broader view and includes all diseases and all possible gene-disease relationships. Furthermore, MedGene utilizes co-citation to indicate a relationship rather than GO annotation, which is limited to the subset of genes that have GO annotation. Our approach is complementary to that taken by Chaussabel and Sher, who used the frequency of co-cited terms to cluster genes into a hierarchy of gene-gene relationships.[6]

A unique aspect of this tool is the ability to assess the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation. This presupposes that most co-citations describe a positive association, often referred to as publication bias[15] and is supported by our observations that negative associations are rare (Supplemental Table 3: http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html). Of course, relationships established by frequency of co-citation do not necessarily represent a true biological link; however, it is strong evidence to support a true relationship.

Another important feature of MedGene is the implementation of software filters that substantially reduced the error rate. We estimate that less than 10% of all associations were missed and at least 70% of even the weakest associations were real. For this study, all of the filters that we applied were general ones, e.g., expanding the list of all gene names to address the different syntax forms used by different journals, eliminating gene names that correspond to common English words, etc. The majority of the remaining search term ambiguities were idiosyncratic and difficult to identify systematically without causing a significant rise in false negatives. Alternative approaches, such as the examination of the nearest neighbor terms, need to be considered to further reduce the false positive rate.

It is not uncommon to see expression changes in microarray experiments as small as 2-fold reported in the literature. Even when these expression changes are statistically significant, it is not always clear if they are biologically meaningful. When comparing expression levels of disease to normal tissue, one expects an enrichment of known disease-related genes to appear in the altered expression group. MedGene provided a unique opportunity to test this notion in the context of existing knowledge on a novel breast cancer micro-array dataset. For genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. This

Table 3. Genes with Large Expression Changes in ER− but Not in ER+ Breast Tumors

gene symbol	fold change (ER+)	fold change (ER−)
KRTHB1	1.0	610.8
BRS3	1.2	89.4
DKK1	1.2	69.8
ZIC1	1.9	59.6
TLR1	1.0	38.5
KIAA0680	2.6	33.2
CDKN3	1.0	30.6
EBI2	4.0	27.9
GZMB	3.8	21.9
STK18	4.7	18.6
GPR49	1.0	14.6
MYO10	1.6	14.4
LAD1	−1.0	13.5
POLE2	4.2	13.0
HMG4	4.4	12.9
BCL2L11	−1.2	12.3
LRP8	2.9	12.2
CCNB2	1.0	11.8
CCNE2	4.0	11.6
FGB	−4.3	11.1
KNSL6	2.9	10.9
H1F5	3.0	10.2
SERPINH2	4.6	10.2
YAP1	1.0	10.0
LPHB	−1.3	−10.4
TCEA2	−1.1	−10.8
TFF1	1.3	−11.4
COL17A1	−4.1	−15.7
POP5	1.1	−16.2
BPAG1	−4.6	−22.3
PDZK1	−1.1	−36.8
VEGFC	−2.8	−51.5
MUC6	−1.4	−64.9
SERPINA5	−1.0	−83.1
MEIS1	−1.6	−85.9
CA12	2.4	−150.3

Table 3. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative, but not ER positive breast tumors. All of these genes have either never been co-cited with breast cancer or have a weak association except those marked with an *.

reflects the many genes whose role in breast cancer may not involve large changes in expression in sporadic tumors (e.g., *BRCA1* and *BRCA2*) and genes whose modest changes in expression may be unrelated to the disease. Strikingly, among genes with a 10-fold change or more in expression level, there was a strong and significant correlation between expression level and a published role in the disease, providing the first global validation of the micro-array approach to identifying disease-specific genes.

The results derived from MedGene have two implications. First, a careful hunt for corroborating evidence of a role in breast cancer should precede any further study of genes with less than 5-fold expression level changes. Second, any genes with 10-fold changes or more are likely to be related to breast cancer and warrant attention. It is likely that this threshold will change depending on the disease as well as the experiment.

Interestingly, the observed correlation was only found among ER-positive tumors, not ER-negative. This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently. The MedGene approach identified a set of relatively understudied, yet highly expressed genes in ER-negative tumors that are worthy of further examination (Table 3).

In conclusion, we have developed an automated method of summarizing and organizing the vast biomedical literature. To our knowledge, the resulting database is the most comprehensive and accurate of its kind. By generating a score that reflects the strength of the association, it provides an important tool for the rapid and flexible analysis of large datasets from various high-throughput screening experiments. Furthermore, it can be used for selecting subsets of genes for functional studies, for building disease-specific arrays, for looking at genes common to multiple diseases and various other high-throughput applications. In the future, it will be possible to enhance the utility of the MedGene database by building links between genes and other MeSH terms as well as other biological processes and concepts, such as cell division and responses to small molecules.

Acknowledgment. We would like to thank P. Braun, L. Garraway, J. Pearlberg, and other members of our institute for helpful discussion. Many thanks to the NLM (National Library of Medicine) for licensing of MEDLINE and the annotation effort of adding MeSH indexes for MEDLINE abstracts. This work was funded by grants from the Breast Cancer Research Foundation and an NHLBI PGA Grant (Vol HL66582-02).

Supporting Information Available: Twenty-three human disease category headings along with all of their child terms selected from the 2002 MeSH index (Supplemental Table 1); analysis of the causes of false negatives in MedGene (Supplemental Table); meaningful gene-disease relationships found in MedGene (Supplemental Table 3); causes for incorrect assignment of gene indexes (Supplemental Table 4); a review of the results, showing that the resulting disease clusters were indeed logical (Supplemental Figure 1); and a review of the results showing that among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease (Supplemental Figure 2). This material is available free of charge via the Internet at http://pubs.acs.org and at the web sites mentioned in the text.

References

(1) Baasiri, R. A.; Glasser, S. R.; Steffen, D. L.; Wheeler, D. A. *Oncogene* **1999**, *18*, 7958−7965.
(2) Steffen, D. L.; Levine, A. E.; Yarus, S.; Baasiri, R. A.; Wheeler, D. A. *Bioinformatics* **2000**, *16*, 639−649.
(3) Marcotte, E. M.; Xenarios, I.; Eisenberg, D. *Bioinformatics* **2001**, *17*, 359−363.
(4) Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. *Bioinformatics* **2001**, *17*, 155−161.
(5) Jenssen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. *Nat. Genet.* **2001**, *28*, 21−28.
(6) Chaussabel, D.; Sher, A. *Genome Biol.* **2002**, *3*, RESEARCH0055.
(7) Gibbons, F. D.; Roth, F. P. *Genome Res.* **2002**, *12*, 1574−1581.
(8) Perez-Iratxeta, C.; Bork, P.; Andrade, M. A. *Nat. Genet.* **2002**, *31*, 316−319.
(9) Funk, M. E.; Reid, C. A. *Bull. Med. Libr. Assoc.* **1983**, *71*, 176−183.
(10) Humphrey, S. M.; Miller, N. E. *J. Am. Soc. Inf. Sci.* **1987**, *38*, 184−196.
(11) Maglott, D. R.; Katz, K. S.; Sicotte, H.; Pruitt, K. D. *Nucleic Acids Res.* **2000**, *28*, 126−128.
(12) Pruitt, K. D.; Maglott, D. R. *Nucleic Acids Res.* **2001**, *29*, 137−140.
(13) Wadelius, M.; Andersson, A. O.; Johansson, J. E.; Wadelius, C.; Rane, E. *Pharmacogenetics* **1999**, *9*, 333−340.
(14) Adam, R. M.; Borer, J. G.; Williams, J.; Eastham, J. A.; Loughlin, K. R.; Freeman, M. R. *Endocrinology* **1999**, *140*, 5866−5875.
(15) Montori, V. M.; Smieja, M.; Guyatt, G. H. *Mayo Clin. Proc.* **2000**, *75*, 1284−1288.
(16) Denenberg, V. H. *Statistics Experimental Design for Behavioral and Biological Researchers*; Wiley-Liss: New York, 1976.
(17) Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J.; Lancet, D. *Trends Genet.* **1997**, *13*, 163.
(18) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45−48.

PR0340227

Genetic Instability in Epithelial Tissues at Risk for Cancer

WALTER N. HITTELMAN

Department of Experimental Therapeutics, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA

ABSTRACT: **Epithelial tumors develop through a multistep process driven by genomic instability frequently associated with etiologic agents such as prolonged tobacco smoke exposure or human papilloma virus (HPV) infection. The purpose of the studies reported here was to examine the nature of genomic instability in epithelial tissues at cancer risk in order to identify tissue genetic biomarkers that might be used to assess an individual's cancer risk and response to chemopreventive intervention. As part of several chemoprevention trials, biopsies were obtained from risk tissues (i.e., bronchial biopsies from chronic smokers, oral or laryngeal biopsies from individuals with premalignancy) and examined for chromosome instability using *in situ* hybridization. Nearly all biopsy specimens show evidence for chromosome instability throughout the exposed tissue. Increased chromosome instability was observed with histologic progression in the normal to tumor transition of head and neck squamous cell carcinomas. Chromosome instability was also seen in premalignant head and neck lesions, and high levels were associated with subsequent tumor development. In bronchial biopsies of current smokers, the level of ongoing chromosome instability correlated with smoking intensity (e.g., packs/day), whereas the chromosome index (average number of chromosome copies per cell) correlated with cumulative tobacco exposure (i.e., pack-years). Spatial chromosome analyses of the epithelium demonstrated multifocal clonal outgrowths. In former smokers, random chromosome instability was reduced; however, clonal populations appeared to persist for many years, perhaps accounting for continued lung cancer risk following smoking cessation.**

KEYWORDS: **chromosome instability; epithelial cells; aerodigestive tract; chemoprevention; cancer risk**

THE NEED FOR BIOMARKERS OF CANCER RISK AND RESPONSE TO INTERVENTION

Epithelial cancers remain a major health challenge in the world. Despite improvements in staging and the application and integration of surgery, radiotherapy, and chemotherapy, the 5-year survival rate for individuals with lung cancer is only about 15%.[1] Even if strategies for early detection are successful and lung cancers are detected at a stage where local tumor resection and treatment is curative, these patients will still be at significant risk for developing second primary tumors

Address for correspondence: Dr. Walter N. Hittelman, Department of Experimental Therapeutics, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. (Box 19), Houston, Texas 77030. Voice: 713-792-2961; fax: 713-792-3754.
whittelm@mdanderson.org

associated with the problem of field cancerization.[2] Similarly, for individuals with a first head and neck primary tumor, even if the first malignancy is successfully treated, the risk of developing a second primary in the tobacco smoke-exposed field is approximately 40%.[3] Similar cancer risk estimates exist for individuals who exhibit severe dysplasia in premalignant epithelial lesions.[4] For these reasons, it is important to focus on chemopreventive strategies to prevent the development of epithelial malignancies.

Several problems confront chemoprevention trials designed to identify efficacious agents.[5] First, chemoprevention trials with cancer incidence as a primary endpoint require tens of thousands of subjects and tens of years of intervention and follow-up for statistical evaluation. For example, a recently reported trial involved 30,000 subjects and required 10 years in order to examine the impact of prevention strategies on lung cancer development, only to find a possible increased lung cancer incidence in current smokers who received β-carotene.[6]

The problem of large, long-term trials results from the difficulty in identifying individuals at highest cancer risk who might best benefit from chemopreventive intervention. For example, 20 pack-year smokers, while known to be at relatively increased risk for developing lung cancer, have approximately a 10% lifetime risk for developing lung cancer.[7] This seriously limits the number of potentially useful strategies that can be clinically explored. A second problem facing chemoprevention trials is that little is known about what agents are likely to have efficacy, and even less is known regarding proper doses, schedules, and durations of treatment. Part of the reason for this problem is that too little is known about the physiologic processes that drive epithelial cancer development.

In order to reduce the number of subjects and the time required to carry out chemoprevention trials and thus allow the exploration of multiple prevention strategies, two types of advances are necessary. First, it is important to identify individuals at significantly increased cancer risk who might best benefit from different types of intervention. Second, in order to allow the rapid identification of agents, doses, and schedules of potentially efficacious agents, it is necessary to identify and validate surrogate endpoints of response that indicate whether the agents are having a positive impact on the target tissue during the chemopreventive intervention.

One approach to identifying individuals at increased aerodigestive tract cancer risk is to explore epidemiologic features of potential subjects. Molecular epidemiologic studies are beginning to identify intrinsic host factors that place some individuals at increased cancer risk, especially those with a chronic smoking history.[8] Most intrinsic factors identified thus far reflect levels of carcinogen metabolism, repair capabilities of the host following DNA damage, and other measures of intrinsic cellular sensitivity to mutagens. While these factors can provide statistically significant risk ratios in case-control studies that are controlled for tobacco exposure, the detected risk ratios usually fall in the range of 1.5 to 10. Unfortunately, this is not sufficient for the individualization of treatment and is not sufficiently high to significantly reduce the numbers of subjects required for chemoprevention trials with cancer incidence as the primary endpoint.

Another approach to identifying individuals at increased cancer risk is to directly examine the target tissue of individuals with known carcinogen exposure (e.g., chronic tobacco smoke exposure), who have evidence of target organ dysfunction

(e.g., chronic obstructive pulmonary disease, changes in voice quality), or who have clinical evidence of premalignancy (e.g., bronchial metaplasia/dysplasia, oral leukoplakia/erythroplakia, cervical intraepithelial neoplasia). The conventional standard for assessing cancer risk in these situations is the degree of histological change. However, while individuals who show moderate to severe dysplasia are known to be at increased cancer risk when compared to individuals with lesser histologic changes, it is often difficult to distinguish reactive changes to carcinogenic insult from initiated and progressing lesions. Similarly, upon cessation of carcinogenic insult, histologic changes may reverse yet cancer risk may continue for many years. For example, while smoking cessation is associated with decreased bronchial metaplasia,[9] increased lung cancer risk continues for many years beyond smoking cessation.[10] In fact, nearly half the newly diagnosed lung cancer cases in the USA occur in former smokers.[11]

The development of assays to identify individuals at high epithelial cancer risk and to directly assess response to intervention in the target tissue is therefore an important research goal. Such assays should be objective and easily quantifiable and, if possible, minimally invasive. Moreover, they should reflect both the disease process and the targeted pathway and thereby be useful in assessing risk and monitoring response to intervention as well as directly testing the hypothesized mechanism of action of the chemopreventive strategy.

In the chemoprevention setting it is important to recognize that one does not know the location of the future cancer. Thus, assays must necessarily be carried out on random biopsies of the field at risk. Even if there are clinically evident premalignant lesions, this does not mean that this is the likely site for a future malignancy. For example, nearly half of the cancers that develop in individuals with oral leukoplakia arise away from the original index lesion. Similarly, since many newly diagnosed lung cancers arise in the peripheral parts of the lung (e.g., adenocarcinomas), especially in former smokers, and since endobronchoscopy predominantly accesses central components of the lung, it is important to identify biomarkers that can reflect global processes ongoing in the target epithelial field associated with increased cancer risk. Their discovery requires a better understanding of the tumorigenesis process in epithelial fields at cancer risk.

THE RATIONALE FOR STUDYING GENOMIC INSTABILITY AS A MARKER OF RISK

Tumors of the aerodigestive tract have been proposed to reflect a "field cancerization" process whereby the whole tissue is exposed to carcinogenic insult (e.g., tobacco smoke) and is at increased risk for multistep tumor development.[12,13] Several types of clinical and laboratory data support this notion, including the frequent occurrence of synchronous primary and subsequent second primary tumors in the aerodigestive tract (frequently exhibiting dissimilar histologies as well as distinct genetic signatures[14–16]) and the presence of premalignant lesions that precede and/or accompany the tumor in the exposed tissue field.[17] The notion of a multistep tumorigenesis process is further supported by serial clinical and histologic evaluations of

target tissue or exfoliated cells where increasing degrees of histological abnormalities are observed over time.[18]

A working model for aerodigestive tract tumorigenesis is illustrated in FIGURE 1. Tumorigenesis in the face of carcinogenic exposure likely involves a chronic process of tissue injury and wound healing. DNA damage induced by the carcinogen is likely fixed into permanent genetic changes (e.g., chromosome damage, chromosome nondisjunction, gene mutation, gene deletion, etc.) during the process of proliferation. This damage would be expected to be distributed throughout the exposed tissue field leading to a background of generalized genomic damage (depicted in FIGURE 1 as a background mat of increasing density). Chronic injury and repair likely leads to the accumulation of cells with increasing amounts of genetic changes as well as the outgrowth of abnormal clones (triangles in FIGURE 1) carrying an accumulation of genetic changes important for selective survival, dysregulated growth, and preferential epithelial take-over by initiated clones (see FIGURE 2).

Cellular and molecular evidence for the field carcinogenesis and multistep tumorigenesis model comes from many laboratories.[19,20] With the advent of a wide array of molecular technologies, a large number of specific molecular genetic and epigenetic changes involving specific oncogenes, tumor suppressor genes, cell regulatory genes, and repair genes have now been described for aerodigestive tract cancers. The identification of these specific molecular changes have now provided probes to explore specific events occurring in premalignant lesions adjacent to aerodigestive tract tumors.[21–24] Frequently, these premalignant lesions showed a subset of the same molecular changes found in the associated tumor, suggesting that these lesions might represent precursor lesions for the associated tumors (i.e., a manifestation of

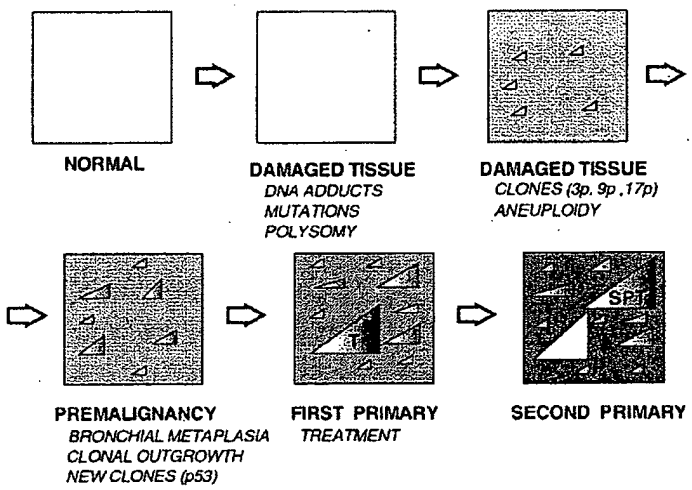


FIGURE 1. Field cancerization and multistep tumorigenesis.

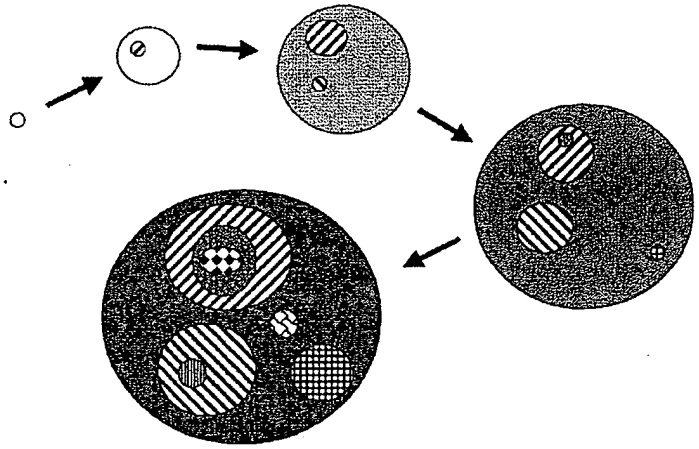


FIGURE 2. Multiple focal clonal evolution during multistep tumorigenesis.

a multistep tumorigenesis process). For example, studies of the premalignant lesions adjacent to head and neck tumors have provided evidence for a gradual accumulation of genetic alterations accompanied by evidence for dysregulation of cellular control mechanisms (e.g., alterations in expression of PCNA, EGFR, TGF-β, p53, and cyclin D1).[25–28]

These types of studies have now also been applied to the target epithelium of individuals at increased risk for aerodigestive tract cancer (i.e., individuals with a chronic smoking/alcohol history and/or prior aerodigestive tract cancer). Several groups (using polymerase chain reaction, PCR, analysis of microdissected epithelium) have now demonstrated the presence of clonal outgrowths in the target premalignant epithelium of individuals at increased risk for cancer.[29–31] For example, examination of bronchial biopsies derived from individuals with a 20 pack-year smoking history demonstrated that 76% of the cases showed evidence for LOH (3p14, 9p21, or 17p13) in at least one of six lung biopsy sites. On a per site basis, some form of LOH was observed in 25% of the sites examined.[29]

If aerodigestive tract cancer development reflects a field cancerization process involving multistep events, then risk and response information should be able to be derived from random biopsies or exfoliated cells from the field at risk or from assessments of tissue undergoing similar processes. Hypothetically, lesions exhibiting the greatest degree of genomic instability, clonal outgrowth, and abnormal epithelial regulation would be at the highest relative aerodigestive tract cancer risk. Similarly, an active chemopreventive intervention might be expected to decrease these manifestations of risk. Reduced risk manifestations include decreased levels of ongoing genetic instability, decreased frequency of clonal outgrowths, and increased epithelial growth regulation.

THE MEASUREMENT OF CHROMOSOME INSTABILITY USING CHROMOSOME *IN SITU* HYBRIDIZATION

Molecular genetic techniques, while extremely useful for detecting clonal changes in targets tissues, are somewhat limited in their ability to detect random genetic instability. Conventional cytogenetic assays are useful for detecting chromosome instability and clonal chromosome changes. However, they require numbers of dividing cells for karyotypic analysis that are difficult to attain in the setting of biopsies acquired during the course of a chemoprevention trial. A technique was therefore needed that would allow chromosome instability measurements in situations where few cells are available (e.g. small biopsies, brushings, or sputum samples) and where the target material might be fixed. It was also desirable to have a technique that would be adaptable to tissue sections, whereby spatial information could be retained and genotype/phenotype associations could be determined on the same or adjacent tissue sections. The technique of *in situ* hybridization (ISH) involves the use of DNA probes that recognize either chromosome-specific repetitive target sequences, chromosome single gene copy sequences, or sequences along the whole chromosome length or chromosome segments.[32] We have adapted the ISH technique for formalin-fixed, paraffin-embedded tissue sections and have applied it to a variety of tissues, including the aerodigestive tract.[33,34]

Using probes that label the centromere regions of specific chromosomes, this assay permits determination of the average chromosome number per cell for each specimen. This assay is also useful for detecting generalized chromosome instability during the tumorigenesis process. Normal diploid populations should have two copies of each autosomal chromosome and should rarely show three or more chromosome copies per cell (chromosome polysomy), especially in tissue sections where nuclear truncation results in an under-representation of chromosome copy number. Thus, the detection of cells with three or more chromosome copies would indicate the presence of chromosome instability.

To examine this technique's potential for characterizing the multistep tumorigenesis process in the aerodigestive tract, we measured the fraction of cells exhibiting three or more chromosome copies in apparently contiguous epithelial transitions from normal to hyperplastic to dysplastic to carcinomas, all on a single tissue slice of head and neck squamous cell carcinomas.[34] In these specimens, greater than 35% of the cases of adjacent "normal" epithelium, greater than 65% of the cases of hyperplastic epithelium, and greater than 95% of the dysplastic and tumor regions showed evidence of chromosome polysomy. Of interest, similar transitions of chromosome instability were observed with at least four different chromosome probes. Similar trends have also been observed in amenable tissue from other epithelial malignancies, including cervix, bladder, and breast.[35] These results thus suggested that the notions of field cancerization and multistep tumorigenesis might apply to several epithelial tissues and that measures of chromosome instability might be useful for monitoring this process.

In the situations described above, the premalignant lesions examined might be considered to represent epithelium at 100% risk of being in a cancer field, since they were located in the adjacent epithelium to the cancer. This then raises the question of the nature of genetic instability in the epithelium of individuals at increased risk

for developing cancer. To explore this issue, we obtained biopsies during the course of leukoplakia chemoprevention trials exploring the use of 13-*cis*-retinoic acid in reversing leukoplakia and probed them for genetic instability using *in situ* hybridization. In one retrospective study and in one prospective study of subjects with oral leukoplakia, the results indicate that those subjects whose pretreatment biopsies harbor relatively high levels of genomic instability (i.e., more than 3% of the cells examined showing at least 3 chromosome 9 copies per cell) have a significantly higher likelihood of suffering early onset of head and neck cancer.[36,37] Interestingly, half of the tumors that did develop occurred away from the biopsy site used to measure genetic instability. This result suggests that genomic instability measurements in carcinogen-exposed tissue can provide useful cancer risk estimates.

THE RELATIONSHIP BETWEEN TOBACCO EXPOSURE AND CHROMOSOME INSTABILITY

In recent years, the aerodigestive tract chemoprevention group at M.D. Anderson Cancer Center has initiated three sequential biomarker-associated chemoprevention trials involving chronic smokers with a greater than 20 pack-year smoking history. In each of these studies, endobronchial biopsies were obtained from six defined sites within the lung, including the carina and at bifurcation points at the upper, middle, and lower right lung and at the upper and lower left lung. Biopsies were obtained prior to and following chemopreventive intervention and were subjected to *in situ* hybridization analysis in addition to analyses for other biomarkers. The first important finding was that some degree of chromosome polysomy was evident in all lung sites examined, and this was observed independently of the particular chromosome probe utilized.[38] This finding supports the notion that random chromosome changes may be occurring throughout the exposed lung field.

In a second study, bronchial biopsies were obtained from individuals with a 20 pack-year smoking history. In this study, most of the subjects involved were current smokers.[39] Interestingly, all cases who showed metaplasia at one of six biopsy sites also showed chromosome polysomy in at least one biopsy site; overall, 88% of the sites showed some evidence of chromosome 9 polysomy.[40] Evidence for genetic instability was also detected in patients who did not show evidence of bronchial metaplasia in any of six biopsy sites despite a strong smoking history. In fact, more than 90% of the cases and more than 60% of the sites showed significant chromosome polysomy (i.e., at least three copies in at least 2 % of the cells examined). These results suggest that the lungs of long-term smokers show significant evidence of genetic instability, and this instability can be detected throughout the accessible bronchial tree, even when bronchial metaplasia is not evident.

These studies in current smokers has allowed us to examine the relationship between the levels of genetic instability detected and subject characteristics such as smoking status (current or former), smoking history, and lung tissue pathologic changes. Evaluable biopsy material has now been obtained from more than 108 current smokers, including more than 480 evaluable biopsy sites. The mean metaplasia index in these current smokers was 30.4%. For the total population studied, the median chromosome index for the bronchial biopsies was 1.41 (range, 1.04–1.61)

and the median chromosome polysomy index was 2.0% (range 0–8.7%). This can be compared to a mean chromosome index between 1.2–1.4 for lymphocytes and very rare chromosome polysomy. Interestingly, the intrasubject variability in chromosome instability was relatively low in most subjects and was less than the intersubject variability. These results suggested that chronic smokers harbor detectable chromosome instability throughout the accessible bronchial tree (supporting the field carcinogenesis notion) and that information from one biopsy site might yield representative information for the rest of the lung field.

Since most of the current smokers exhibited bronchial metaplasia in at least one of the biopsied sites, this allowed us to examine the relationship between chromosome instability and histologic changes, both on a site-by-site basis and on a per case basis. On a site-by-site basis, the chromosome indices of lesions showing squamous metaplasia were similar to those not showing metaplasia (i.e., median 1.43 vs. 1.43), and the degree of chromosome polysomy in metaplastic lesions were only slightly higher than in non-metaplastic sites (medians: 2.2% vs. 1.8%, respectively). Thus, the presence or absence of squamous metaplasia at a biopsy site does not necessarily correlate with the degree of underlying genomic instability. On the other hand, those subjects with metaplasia indices of at least 15% also showed higher levels of chromosome polysomy than did subjects with metaplasia index below 15% (medians: 2.4% vs. 1.8%, $p = 0.005$). Thus, these chromosome instability assessments in current smokers appeared to reflect a more global process in the lung field.

Tobacco exposure has been shown to significantly increase the risk of developing lung cancer, and the degree of risk is related to the extent of tobacco exposure. We were interested in determining the relationship between individuals' smoking history parameters and the levels of chromosome change found in their lungs following years of tobacco exposure. While there was significant intersubject variation for similar tobacco exposure histories, overall there was a significant correlation between the degree of chromosome polysomy and the intensity of ongoing tobacco exposure (packs/day, $p = 0.02$ on a per site basis) and with the extent of tobacco exposure (pack-years, $p = 0.003$). Thus the amount of chromosome polysomy reflects the intensity and extent of tobacco exposure. At the same time, individuals with similar smoking histories showed widely divergent amounts of chromosome polysomy, possibly reflecting differences in intrinsic sensitivity between subjects. There was also strong correlation between the chromosome index and the duration of the smoking history (smoking years) and total accumulated exposure (pack-years, $p = 0.0001$). These results suggest that tobacco exposure is associated with the initiation and accumulation of chromosome instability in the exposed lung; however individuals are differentially sensitive to carcinogenic insult. The working hypothesis is that those individuals who accumulate the highest degree of chromosome changes will be at the highest lung cancer risk.

Many of the bronchial biopsies from chronic smokers examined by *in situ* hybridization showed a rise in the chromosome index above that expected for a diploid cell population, especially in subjects with an extensive smoking history. The rise in chromosome index was also accompanied by an increase in the fraction of cells exhibiting at least 3 chromosome copies per cell. To determine if a rise in the tissue chromosome index was due to clonal expansion of populations with chromosome trisomy, the chromosome copy number and relative coordinates of each cell scored in

the bronchial epithelium was recorded and a spatial genetic map was created.[41] We then developed algorithms for calculating localized chromosome indices within the tissue. Since trisomic clones would have, on average, three chromosomes instead of two, those cells involved in neighborhoods with chromosome indices three-halves that of diploid populations could be marked as being part of a trisomic clone. Similarly, groups of cells with chromosome indices half that of diploid populations could be marked as being part of a monosomic clone. This allowed the generation of a second-order, two-dimensional genetic map representation of the bronchial epithelium showing the relative locations of cells involved in monosomic and trisomic clonal outgrowths. When adjacent tissue sections from the same bronchial biopsy were probed separately for different chromosomes, the detected clones appeared to occupy separate subregions of the epithelium. This result suggests that not only are the lungs of chronic smokers undergoing a process of genetic instability, they are experiencing the outgrowth of multiple clones throughout the exposed lung field, as postulated by the models shown in FIGURES 1 and 2. One advantage of this clonal approach is that the contribution of both monosomic and multisomic clones can be detected.

Since smoking cessation has been suggested to reduce the lung cancer risk, it was of interest to determine whether the levels of chromosome instability would decrease following smoking cessation. This question was possible to examine because our third sequential chemoprevention trial involved subjects who had discontinued smoking. So far, more than 220 subjects (more than 650 biopsies) who have quit smoking (mean 9.9 quit-years) have been evaluated for chromosome instability in their lungs. Despite the fact that the mean metaplasia index in this group is 5.8% (considerably less than that in current smokers), chromosome instability is still observed in the majority of subjects.[42] While the mean chromosome polysomy level is reduced to 1.0%, some individuals continue to show polysomy levels above 5%. Interestingly, while the overall chromosome polysomy levels were reduced in these individuals who stopped smoking, the mean chromosome index remained at about 1.4 with some individuals exhibiting chromosome indices as high as 1.8. Initial chromosome mapping studies suggest that while random chromosome instability seems to decrease following smoking cessation, the clonal outgrowths may remain for many years in the lung. The working hypothesis is that those individuals who show the greatest degree of remaining chromosome instability are at the highest lung cancer risk despite smoking cessation. Long-term follow-up on these subjects will be necessary to test this hypothesis.

SUMMARY AND CONCLUSIONS

Aerodigestive tract tumorigenesis appears to be a multistep process taking place throughout the tissue fields of exposure. When viewed in the context of chromosome changes, carcinogen exposure appears to be associated with the random acquisition of chromosome polysomy throughout the exposed field, the degree of which is related to the degree and extent of carcinogen exposure as well as to the instrinsic susceptibility of the exposed individual. Continued exposure leads to continued acquisition of new changes and, in association with chronic wound-healing processes, to the

accumulation of clonal outgrowths throughout the target tissue. Although the ultimate malignancy may occur in only one or few tissue sites, manifestations of the instability process that drives tumorigenesis is globally present in the tissue. Thus random biopsies may provide useful risk information for the exposed field as a whole. Even when carcinogen exposure is reduced or chemopreventive strategies are initiated and histologic manifestations of the tumorigenesis process subside, the genetic scars of prior exposure remain in the form of clonal outgrowths and may explain continued lung cancer risk in ex-smokers. Future chemoprevention strategies need to focus on reducing the degree of chromosome instability and on trying to eliminate residual abnormal clonal outgrowths in the aerodigestive tract. In this setting, the measurement of chromosome instability in the target tissue will be useful in assessing cancer risk as well as response to intervention.

ACKNOWLEDGMENTS

The studies reviewed here represent one component of the collaborative efforts of the Aerodigestive Tract Chemoprevention team at The University of Texas M.D. Anderson Cancer Center, Houston, Texas. The studies were supported in part by National Institutes of Health-National Cancer Institute Grants CA 52051, CA 68437, CA 79437, CA 16672, CA 68089, CN 25433, CA 86390, CA 70907, NIH DE 13157, and the State of Texas Tobacco Research Fund.

REFERENCES

1. Landis, S.H . T. Murray, S. Bolden & P.A. Wingo. 1998. Cancer statistics, 1998. CA Cancer J. Clin. **48:** 6–29.
2. Johnson, B.E. 1998. Second lung cancers in patients after treatment for an initial lung cancer. J Natl. Cancer Inst. **90:** 1335–1345.
3. Lippman, S.M. & W.K. Hong. 1989. Second malignant tumors in head and neck squamous cell carcinoma. The overshadowing threat for patients with early stage of disease. Int. J. Radiat. Oncol. Biol. Phys. **17:** 691–694.
4. Silverman, S.J., Jr., M. Gorsky & F. Lozada. 1984. Oral leukoplakia and malignant transformation: a follow-up study of 257 patients. Cancer **53:** 563–568.
5. Lippman, S.M., J.S. Lee, R. Lotan, *et al.* 1990. Biomarkers as intermediate endpoints in chemoprevention trials. J. Natl. Cancer Inst. **82:** 555–560.
6. Heinonen, O.P., D. Albanes & The Alpha-tocopherol, Beta Carotene Cancer Prevention Study Group. 1994. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N. Engl. J. Med. **330:** 1029–1035.
7. Peto, R., S. Darby, H. Deo, *et al.* 2000. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Brit. Med. J. **321:** 323–329.
8. Perera, F.P. 1996 Molecular epidemiology: insights into cancer susceptibility, risk assessment, and prevention. J. Natl. Cancer Inst. **88:** 496–509.
9. Lee, J.S., S.M. Lippman, S.E. Benner, *et al.* 1994. Randomized placebo-controlled trial of isotretinoin in chemoprevention of bronchial squamous metaplasia. J. Clin. Oncol. **12:** 937–941.

10. U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. 1990. The health benefits of smoking cessation: a report of the Surgeon General. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. DHHS Pub. No. (CDC) 90-8416.
11. TONG, L., M.R. SPITZ, J.J. FAEGER, *et al.* 1996. Lung cancer in former smokers. Cancer **78:** 1004–1010.
12. SLAUGHTER, D.P., H.W. SOUTHWICK & W. SMEJKAL. 1953. Field cancerization in oral stratified squamous epithelium: clinical implications of multicentric origin. Cancer **6:** 963–968.
13. FARBER, E. 1984. The multistep nature of cancer development. Cancer Res. **44:** 4217–4223.
14. CHUNG, K.Y., T. MUKHOPADHYAY, J. KIM, *et al.* 1993. Discordant p53 gene mutations in primary head and neck cancers and corresponding second primary cancers of the upper aerodigestive tract. Cancer Res. **53:** 1676–1683.
15. SCHOLES, A.G.M., J.A. WOOLGAR, M.A. BOYLE, *et al.* 1998. Synchronous oral carcinomas: independent or common clonal origin? Cancer Res. **58:** 2003–2006.
16. GLUCKMAN, J.O., J.D. CRISSMAN & J.O. DONEGAN. 1980. Multicentric squamous cell carcinoma of the upper aerodigestive tract. Head Neck Surg. **3:** 90–96.
17. AUERBACH, O., A.P. STOUT, E.C. HAMMOND, *et al.* 1961. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. N. Engl. J. Med. **265:** 253–267.
18. SACCOMANNO, G., V.E. ARCHER, O. AUERBACH, *et al.* 1974. Development of carcinoma of the lung as reflected in exfoliated cells. Cancer **33:** 256–270.
19. IZZO, J.G. & W.N. HITTELMAN. 1999. Characterization of multistep tumorigenesis by in situ hybridization. *In* Introduction to Fluorescence In Situ Hybridization: Principles and Clinical Applications. M. Andreeff & D. Pinkel, Eds.: 173–208. John Wiley & Sons, Inc. New York.
20. HITTELMAN, W.N. 1999. Molecular cytogenetic evidence for multistep tumorigenesis: implications for risk assessment and early detection. *In* Molecular Pathology of Cancer. S. Srivastava, D.E. Hensen & A. Gazdar, Eds.: 385–404. IOS Press. Amsterdam, The Netherlands.
21. SUNDARESAN, V., P. GANLY, R. HASLETON, *et al.* 1992. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumours, are detectable in preinvasive lesions of the bronchus. Oncogene **7:** 1989–1997.
22. KISHIMOTO, Y., K. SUGIO, J.Y. HUNG, *et al.* 1995. Allele-specific loss in chromosome 9p loci in preneoplastic lesions accompanying non-small-cell lung cancers. J. Natl. Cancer Inst. **87:** 1224–1229.
23. CALIFANO, J., P. VAN DER RIET, W. WESTRA, *et al.* 1996. Genetic progression model for head and neck cancer: implications for field cancerization. Cancer Res. **56:** 2488–2492.
24. PARK I.W., I.I. WISTUBA, A. MAITRA, *et al.* 1999. Multiple clonal abnormalities in the bronchial epithelium of patients with lung cancer. J. Natl. Cancer Inst. **91:** 1863–1868.
25. SHIN, D.M., N. VORAVUD, J.Y. RO, *et al.* 1994. Sequential increases in proliferating cell nuclear antigen expression in head and neck tumorigenesis: a potential biomarker. J. Natl. Cancer Inst. **85:** 971–978.
26. SHIN, D.M., J.Y. RO, W.K. HONG, *et al.* 1994. Dysregulation of epidermal growth factor receptor expression in premalignant lesions during head and neck tumorigenesis. Cancer Res. **54:** 3153–3159.
27. SHIN, D.M., J. KIM, J.Y. RO, *et al.* 1994. Activation of p53 gene expression in premalignant lesions during head and neck tumorigenesis. Cancer Res. **54:** 321–326.
28. IZZO, J.G., V.A. PAPADIMITRAKOPOULOU, X.Q. LI, *et al.* 1998. Dysregulated cyclin D1 expression early in head and neck tumorigenesis: in vivo evidence for an association with subsequent gene amplification. Oncogene **17:** 2313–2322.
29. MAO, L., J.S. LEE, J.M. KURIE, *et al.* 1997. Clonal genetic alterations in the lungs of current and former smokers. J. Natl. Cancer Inst. **89:** 857–862.

30. WISTUBA, I.I., S. LAM, C. BEHRENS, *et al.* 1997. Molecular damage in the bronchial epithelium of current and former smokers. J. Natl. Cancer Inst. **89:** 1366–1373.
31. MAO, L., J.S. LEE, Y.H. FAN, *et al.* 1996. Frequent microsatellite alterations at chromosomes 9p21 and 3p14 in oral premalignant lesions and their value in cancer risk assessment. Nature Med. **2:** 682–685.
32. PODDIGHE, P.J., F.C. RAMAEKERS & A.H. HOPMAN. 1992. Interphase cytogenetics of tumours. J. Pathol. **166:** 215–224.
33. KIM, S.Y., J.S. LEE, J.Y. RO, *et al.* 1993. Interphase cytogenetics in paraffin sections of lung tumors by non-isotopic in situ hybridization. Mapping genotype/phenotype heterogeneity. Am. J. Pathol. **142:** 307–317.
34. VORAVUD, N., D.M. SHIN, J.Y. RO, *et al.* 1993. Increased polysomies of chromosomes 7 and 17 during head and neck multistage tumorigenesis. Cancer Res. **53:** 2874–2883.
35. HITTELMAN, W.N. 1999. Genetic instability assessments in the lung cancerization field. *In* Lung Tumors: Fundamental Biology and Clinical Management. C. Brambilla & E. Brambilla, Eds.: 255–267. Marcel Dekker. New York.
36. LEE, J.S., S.Y. KIM, W.K. HONG, *et al.* 1993. Detection of chromosomal polysomy in oral leukoplakia, a premalignant lesion. J. Natl. Cancer Inst. **85:** 1951–1954.
37. LEE, J.J., W.K. HONG, W.N., HITTELMAN, *et al.* 2000. Predicting cancer development in oral leukoplakia: ten years of translational research. Clin. Cancer Res. **6:** 1702–1710.
38. HITTELMAN W.N., R. YU, J. KURIE, *et al.* 1997. Evidence for genomic instability and clonal outgrowth in the bronchial epithelium of smokers [abstract]. Proc. Am. Assoc. Cancer Res. **38:** 3097.
39. KURIE, J.M., J.S. LEE, F.R. KHURI, *et al.* N-(4-hydroxyphenyl)retinamide in the chemoprevention of squamous metaplasia and dysplasia of the bronchial epithelium. 2000. Clin. Cancer Res. **6:** 2973–2979.
40. HITTELMAN, W.N., J.S. LEE, R.C. MORICE, *et al.* 1999. Lack of biomarker modulation in bronchial biopsies of chronic smokers following treatment with N-(4-hydroxyphenyl)retinamide (4-HPR). Proc. Am. Assoc. Cancer Res. **40:** 2837.
41. HITTELMAN, W.N., J.S. LEE, N. CHEONG, *et al.* 1991. The chromosome view of "field cancerization" and multistep carcinogenesis. Implications for chemopreventive approaches. *In* Chemoimmunoprevention of Cancer. V. Pastorino & W.K. Hong, Eds.: 41–47. Georg Thieme Verlag. Stuttgart, Germany.
42. HITTELMAN, W.N., J.J. LEE, J.S. LEE, *et al.* 1998. Persistent genetic instability despite decreased proliferation in human lung tissue following smoking cessation. Proc. AACR **39:** 336.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 277, No. 35, Issue of August 30, pp. 31291–31302, 2002
Printed in U.S.A.

A Genomic and Proteomic Analysis of Activation of the Human Neutrophil by Lipopolysaccharide and Its Mediation by p38 Mitogen-activated Protein Kinase*

Received for publication, January 24, 2002
Published, JBC Papers in Press, April 9, 2002, DOI 10.1074/jbc.M200755200

Michael B. Fessler‡§, Kenneth C. Malcolm§, Mark William Duncan¶, and G. Scott Worthen‡§‖

From the ‡Department of Medicine, Division of Pulmonary Sciences and Critical Care Medicine, University of Colorado Health Sciences Center, the §Department of Medicine, National Jewish Medical and Research Center, and the ¶Biochemical Mass Spectrometry Facility, School of Pharmacy, University of Colorado Health Sciences Center, Denver, Colorado 80262

Bacterial lipopolysaccharide (LPS) evokes several functional responses in the neutrophil that contribute to innate immunity. Although certain responses, such as adhesion and synthesis of tumor necrosis factor-α, are inhibited by pretreatment with an inhibitor of p38 mitogen-activated protein kinase, others, such as actin assembly, are unaffected. The aim of the present study was to investigate the changes in neutrophil gene transcription and protein expression following lipopolysaccharide exposure and to establish their dependence on p38 signaling. Microarray analysis indicated expression of 13% of the 7070 Affymetrix gene set in nonstimulated neutrophils, and LPS up-regulation of 100 distinct genes, including cytokines and chemokines, signaling molecules, and regulators of transcription. Proteomic analysis yielded a separate list of up-regulated modulators of inflammation, signaling molecules, and cytoskeletal proteins. Poor concordance between mRNA transcript and protein expression changes was noted. Pretreatment with the p38 inhibitor SB203580 attenuated 23% of LPS-regulated genes and 18% of LPS-regulated proteins by ≥40%. This study indicates that p38 plays a selective role in regulation of neutrophil transcripts and proteins following lipopolysaccharide exposure, clarifies that several of the effects of lipopolysaccharide are post-transcriptional and post-translational, and identifies several proteins not previously reported to be involved in the innate immune response.

Lipopolysaccharide (LPS),[1] a component of the outer cell wall of Gram-negative bacteria, evokes a variety of functional responses in the human neutrophil (PMN) after binding to a plasma membrane receptor complex that involves the Toll-like receptors (TLRs) (1–5). These "immediate" functional responses, including actin assembly, adhesion, activation of nuclear factor-kappa B (NF-κB), and priming for an enhanced secretory response and for release of reactive oxygen intermediates, appear to be central both to the innate immune response and to the pathogenesis of several inflammatory human diseases, including sepsis and the acute respiratory distress syndrome (6). p38 mitogen-activated protein kinase (p38 MAPK) has been shown to mediate LPS-induced PMN adhesion, NF-κB activation, and TNF-α and IL-8 translation and release (7), and its blockade attenuates LPS-induced PMN accumulation in the airspace (8). However, other cascades almost certainly lead to downstream effectors of the LPS signal; for example, actin assembly appears to be p38 MAPK-independent (9). An improved understanding of the transcriptional and translational responses of the neutrophil to LPS and the modulation of these responses by p38 MAPK might carry pathogenetic and therapeutic implications.

Historically, it has been believed that the downstream PMN transcriptional response to LPS is static and that PMN functional responses to LPS that depend on *de novo* protein synthesis are primarily limited to the release of cytokines (10). However, recent studies indicate a robust transcriptional response (11). To date, most studies have relied upon and reported a short list of functional assays of the LPS-exposed PMN; therefore, no exhaustive investigation of either the transcriptional response or protein synthetic repertoire of the PMN has been reported. Although several techniques have been used to evaluate transcripts, the screening of global changes in mRNA by microarray analysis has only recently become possible. In this way, thousands of genes can be screened in an unbiased fashion for transcript abundance. Such genomic screens in mammalian cells have previously been applied to define altered expression profiles in response to agonists (12) and to drug action (13) and during cell cycle progression (14).

Although DNA microarray technology is expected to provide insight into the response of the human PMN to LPS (15), inhibition of LPS-stimulated IL-1 and TNF-α production by p38 MAPK inhibitors in THP-1 cells (16) and of TNF-α synthesis in human PMNs (9) occurs at a translational level and would therefore not be detected by DNA microarrays. Furthermore, in other systems, such as yeast and human liver, mRNA and protein levels show poor correlation (17, 18). Proteomics is a complementary tool for assessing global changes in cellular protein expression, thereby providing additional insight into cellular signal regulation. A proteomic approach has proven useful in different systems for dissecting signal transduction cascades and describing their output (19, 20) and has even

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‖ To whom correspondence should be addressed: Dept. of Medicine, D403, Neustadt Bldg., National Jewish Medical and Research Center, 1400 Jackson St., Denver, CO 80206. Tel.: 303-398-1171; Fax: 303-398-1381; E-mail: worthens@njc.org.

[1] The abbreviations used are: LPS, lipopolysaccharide; DTT, dithiothreitol; IEF, isoelectric focusing; IFN, interferon; IL, interleukin; MALDI-TOF, matrix-assisted laser desorption ionization-time of flight; MAPK, mitogen-activated protein kinase; NF-κB, nuclear factor-kappa B; pI, isoelectric point; PMN, neutrophil (polymorphonuclear leukocyte); TLR, Toll-like Receptor; TNF, tumor necrosis factor; CHCA, α-cyano-4-hydroxycinnamic acid; AEBSF, 4-(2-aminoethyl-)benzenesulfonylfluoride hydrochloride; MS, mass spectrometry; CaM, Ca^{2+}/calmodulin; ERK, extracellular signal-regulated kinase; E-64, epoxysuccinyl-64.

recently been used to detect novel upstream messengers involved in LPS signal transduction (21). We have applied DNA microarrays and proteomics to define and compare transcriptional and post-transcriptional alterations in the LPS-exposed PMN and to establish the dependence of these alterations on p38 MAPK signaling.

EXPERIMENTAL PROCEDURES

Materials—Endotoxin-free reagents and plastics were used in all experiments. Aprotinin, leupeptin, AEBSF, E-64, pepstatin, and bestatin protease inhibitors, spermine HCl, and α-cyano-4-hydroxycinnamic acid (CHCA) were all purchased from Sigma Chemical Co. (St. Louis, MO). SB203580, a p38 MAPK inhibitor, was purchased from Calbiochem-Novabiochem Corp. (San Diego, CA). For two-dimensional PAGE, rehydration buffer, equilibration buffers, vertical electrophoresis solutions, and 10% homogeneous polyacrylamide slab gels were purchased from Genomic Solutions, Inc. (GSI, Ann Arbor, MI). Sequencing grade porcine trypsin was purchased from Promega (Madison, WI).

LPS Incubation—PMNs were isolated by the plasma Percoll method (22), a technique that yields less than 5% monocytic contamination, and resuspended at a concentration of 15.4×10^6/ml in RPMI 1640 culture medium (BioWhittaker, Walkersville, MD) supplemented with 10 mM HEPES (pH 7.6) and 1% heat-inactivated platelet-poor plasma. After addition of 100 ng/ml *Escherichia coli* 0111:B4 LPS (List Biological), incubation was carried out with continuous rotation (4 h, 37 °C) both in the presence and absence of SB203580. Both Affymetrix analysis and proteomic analysis utilized 75×10^6 cells. For microarray analysis, nonstimulated and 4-h-treated PMNs were collected from three separate donors. A more detailed time course following LPS exposure was performed using polymerase chain reaction. For proteomic analysis, LPS incubations from separate donors ($n = 6$) were performed and then analyzed individually. Control and post-LPS incubation PMNs were washed (0.34 M sucrose/1 mM EDTA/10 mM Tris) and then lysed in a modified rehydration buffer (GSI, Ann Arbor, MI) supplemented with 2 M thiourea, 50 mM dithiothreitol (DTT), 22.5 mM spermine HCl, and a mixture of six protease inhibitors (10 μg/ml aprotinin, 10 μg/ml leupeptin, 2 mM AEBSF, 5 μM E-64, 1 μM pepstatin, 10 μM bestatin). DNA was pelleted by centrifugation at $250,000 \times g$ for 60 min (23).

Affymetrix Oligonucleotide Array—Five micrograms of total RNA was isolated with TRIzol (Invitrogen) and RNeasy columns (Qiagen) and subsequently labeled with biotin as described by Affymetrix. Briefly, first-strand synthesis was accomplished with Superscript II reverse transcriptase (Invitrogen) using a T7-oligo$(dT)_{24}$ primer for 1 h at 42 °C followed by second-strand synthesis using *E. coli* DNA polymerase I and RNase H (Invitrogen) at 16 °C for 2 h. Double-stranded DNA was used as a template for *in vitro* transcription with T7 RNA polymerase in the presence of biotin-labeled UTP and CTP using the BioArray High Yield RNA transcript labeling kit (Enzo). Fifteen micrograms of cRNA was fragmented and used for hybridization to Affymetrix HuGene 6800FL Genechips. Each sample was hybridized initially using a Test2 Genechip to test for sample degradation and full-length *in vitro* translation. Data were analyzed using Affymetrix Genechip software. Results from three separate donors were analyzed.

Reverse Transcription and Polymerase Chain Reaction—cDNA was prepared by reverse transcription using 2 μg total RNA, derived from 20×10^6 cells that were treated as indicated. Polymerase chain reactions were performed using specific primers for *Mx-1*, *TNF-α*, *MCP-1*, *p65*, *S100A4*, and glyceraldehyde-3-phosphate dehydrogenase.

Two-dimensional PAGE—The protein concentration of the lysates was measured as described by Bradford *et al.* (24). Poor isoelectric focusing (IEF) results were encountered unless the polycationic spermine was diluted (data not shown); therefore, lysates were diluted with rehydration buffer (GSI, Ann Arbor, MI) to achieve a final spermine concentration of 6 mM. Equal protein loads (1.5 mg) of control and LPS-stimulated neutrophils were used to rehydrate IEF gels overnight (18 cm, pH 3–10 nonlinear Immobiline DryStrip IEF gels, Amersham Biosciences; Piscataway, NJ). IEF was performed at 20 °C to 100-kVh (Phaser, GSI) under mineral oil, followed by two 10-min SDS equilibration steps (DTT and then iodoacetamide-containing equilibration buffers, GSI) and then by vertical electrophoresis on 10% homogeneous polyacrylamide slab gels (GSI) at 500 V. Protein spots were visualized by agitation in colloidal Coomassie Brilliant Blue G-250 (16 h) (25), followed by destaining in deionized water (20 h). In separate experiments, control and LPS-stimulated PMN lysates from three donors were pooled and then analyzed by two-dimensional PAGE using overlapping narrow isoelectric point (pI) ranges (18 cm, pH 5.0–6.0, 5.5–6.7, and 6–11, Amersham Biosciences, Piscataway, NJ). Identical IEF and vertical electrophoresis parameters were used for all gels.

Image Analysis of Two-dimensional Gels—Colloidal Coomassie-stained gels were digitized using a Powerlook II (UMAX Data Systems, Inc., Taiwan) flatbed scanner with 8-bit dynamic range and 150-dpi resolution. BioImage (GSI, Ann Arbor, MI) 2D-Analyzer software was used to locate, quantitate, and match protein spots on the control and LPS gel images. Analysis was performed by assigning 50 common anchor spots between paired images; the remaining spots were compared by a constellation-matching algorithm. All data were then carefully reviewed by the operator to account for any discrepancies. Protein loading between control and experimental gels may have varied because of inconsistencies in rehydration of the different IEF gel strips; therefore, gel images were normalized so that the sum of the integrated intensities of all matched spots on paired gels was made equal. Control and LPS-stimulated gel images from individual donor experiments were matched to generate composite images; composite images were then matched into a master composite image to track the LPS response of protein spots among different donors (26). Only those spots that were common (image-matched) to all original 12 (pH 3.0–10.0) gels were considered for further analysis. For these spots, the LPS-induced change in integrated intensity in the six experiments was subjected to statistical analysis with a two-tailed Student's t test, and those spots with $p < 0.05$ were identified by peptide mass fingerprinting (described below). For the narrow range (pH 5.0–6.0, 5.5–6.7, and 6–11) two-dimensional PAGE experiments using pooled donors, only those spots with concordant regulation exceeding 1.5-fold or that appeared *de novo* in the LPS gel in two repeat experiments were further analyzed.

In-gel Tryptic Digestion—In-gel digestion of protein spots was performed with sequencing grade porcine-modified trypsin using the method of Hellman *et al.* (27). Tryptic peptides were then extracted (50 μl of 50% acetonitrile/5% trifluoroacetic acid, 2 h), and the supernatant was taken to dryness in a vacuum centrifuge and then redissolved in trifluoroacetic acid (20 μl, 0.5%). Peptides were then purified and concentrated using ZipTip$_{C18}$ pipette tips (Millipore, Bedford, MA).

MALDI-TOF Mass Spectrometry—Analyses were performed on an Applied Biosystems matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) Voyager-DE PRO mass spectrometer (Framingham, MA) operated in delayed extraction mode. Samples (0.5 μl) were spotted onto a sample plate to which matrix (0.5 μl of 10 mg/ml CHCA) was added. The sample-matrix mixture was dried at room temperature and then analyzed in reflector mode. CHCA was also spotted alone as a negative control. Spectra were the sum of 100 laser shots, and those peaks with a signal-to-noise ratio of greater than 3:1 were selected for data base searching. Spectra were internally calibrated using autolytic trypsin peptides (*m/z* 842.51, 2211.10).

Data Base Searching Algorithm—The monoisotopic masses for each protonated peptide were: (*a*) entered into the program MS-Fit (available at prospector.ucsf.edu) for searches against the Swiss-Prot, NCBI, and GenPept databases, and (*b*) entered into Mascot (available at matrixscience.com), an algorithm testing statistical significance of peptide mass fingerprinting identifications. For MS-Fit searches, masses derived from trypsin, CHCA, keratin, and Coomassie Brilliant Blue G-250 were excluded. Search parameters included a maximum allowed peptide mass error of 0.1 Da (0.8 Da in the few instances in which linear mode was used), consideration of one incomplete cleavage per peptide, pI range of 3.0–10.0, and molecular mass range of 1–200 kDa. Accepted modifications included carbamidomethylation of cysteine residues (from iodoacetamide exposure following IEF) (28) and methionine oxidation, a common modification occurring during SDS-PAGE (29). Protein identifications were assigned when three criteria were met: 1) statistical significance ($p < 0.05$) of the match when tested by Mascot (matrixscience.com); 2) >20% sequence coverage by the tryptic peptides; and 3) concordance (±15%) with the molecular weight and pI of the parent two-dimensional PAGE protein spot. The following special exceptions were considered: (*a*) protein identifications not fulfilling criterion 2 were still assigned if criteria 1 and 3 were fulfilled and no other *Homo sapiens* proteins with peptide mass-matched p values < 0.05 were identified by Mascot; (*b*) if criterion 3 was not fulfilled (lower than expected molecular weight), a cleavage product of the identified protein was inferred, and the cumulative molecular weight of the tryptic peptides was compared with that of the two-dimensional-PAGE spot to ensure that it was not exceeded; (*c*) if criterion 3 was not fulfilled (isolated discordance between theoretical and observed pI), post-translational modification of an unrecovered peptide was inferred; and (*d*) if two or more *H. sapiens* protein assignments with >4 mutually exclusive matching peptides were identified, a protein mixture in the two-dimensional PAGE

spot was inferred and further analysis halted (quantitative conclusions regarding the individual protein constituents could not be drawn).

RESULTS

Genes Differentially Expressed in LPS-stimulated Neutrophils—Human PMNs were left untreated or incubated in the presence of 100 ng/ml LPS for 4 h. As a control to confirm that the PMNs were quiescent at baseline and that LPS resulted in normal stimulation, mRNA was isolated, cDNA was prepared, and PCR for TNF-α was performed. Little TNF-α expression was seen in nonstimulated cells, whereas LPS treatment led to an increase in expression in each of the donors subsequently used for microarray analysis (data not shown). No macrophage-colony stimulating factor receptor transcript was detected by oligonucleotide microarray analysis, confirming there was no significant monocytic contamination.

Human PMNs express a limited repertoire of mRNA transcripts at baseline but respond to LPS with differential expression of genes in many families. Considering only those genes present by microarray analysis in all three donors, unstimulated PMNs expressed 13.0% (923 of 7070 genes) of the Affymetrix gene set. Gene classes represented at baseline include metabolic enzymes, structural proteins, receptors, signaling proteins, and transcription factors. By comparison, human monocytes expressed ~40% and human fibroblasts ~35% of the represented genes (data not shown). By the criterion of a >3-fold increase in expression in all three donors on Affymetrix oligonucleotide array analysis, exposure of PMNs to LPS for 4 h resulted in the up-regulation of 100 genes (Table I).

Genes from several different functional classes were induced in PMNs following LPS exposure. Of interest, a number of transcriptional regulators were induced, including transcription factors of the NF-κB family. The transcriptional NF-κB complex has previously been implicated in the regulation of the genes induced by LPS (11). The genes for several cytokines and chemokines were also found to be up-regulated. These include *TNF-α*, *IL-1β*, *IL-6*, *MCP-1*, *MIP-3α*, and *MIP-1β* (Table I). PCR was performed to confirm the results from the microarray analysis. PCR analysis on selected genes indicates that the time course for changes can be rapid or delayed but parallel the changes found in the array at the 4-h time point (data not shown). Other up-regulated genes included those for metabolic enzymes, immune response molecules, kinases, phosphatases, signaling molecules, adhesion and cytoskeletal components, interferon-stimulated genes, and those with unknown or miscellaneous function (Table I).

LPS stimulation of PMN also resulted in the down-regulation of 56 genes (Table II). Down-regulated genes were identified as transcriptional regulators, protein and lipid kinases and phosphatases, structural molecules, and signaling molecules. Genes for metabolic proteins were also evident, as were several uncharacterized genes.

Two-dimensional PAGE and Image Analysis—In contrast to the limited number of transcripts found at baseline, PMNs were found to express a large number and variety of proteins in the nonstimulated state (Fig. 1, *A* and *C*, and Tables III–V). Reproducible protein expression patterns were found on the pH 3.0–10.0 gels, and the majority of proteins fell in the pH 5.0–7.0 range (Fig. 1*A*). The basic region (pH > 7.0) consistently exhibited poor resolution, precluding meaningful image analysis and further workup (data not shown). Depending on the spot-finding parameters (minimum spot intensity, filter width) selected on the image analysis software, spot-by-spot manual editing was found to be necessary to avoid over- and underdetected spots; moreover, further manual editing was performed to screen for unmatched and mismatched spots following matching of paired control and LPS-stimulated gels. After spot editing, ~1200 well-resolved spots were evident on each pH 3.0–10.0 gel. In an attempt to improve resolution of the pI range bearing the greatest number of well-resolved spots, overlapping narrow pH range gels (pH 5.0–6.0, 5.5–6.7, 6–11) were also run. Of interest, a similar number of well-resolved spots (~1200) were detected on the narrow pH range gels (Fig. 1, *C* and *D*). Assuming a detection limit for Coomassie of 15 ng (0.25 pmol, or 1.5×10^{11} molecules, for a 60-kDa protein) and a protein load per gel corresponding to 75×10^6 PMNs, we estimate a detection limit on our gels of 2000 molecules/cell for a 60-kDa protein. As investigators have suggested in other cell lines with the use of high resolution two-dimensional-PAGE methods (30), we estimate that >10,000 proteins are expressed in the resting PMN.

Human PMNs respond to LPS with the differential expression of a large number of proteins. In the six individual pH 3.0–10.0 experiments, the number of protein spots that increased in integrated intensity by at least 50% following LPS exposure was 185, 122, 104, 104, 96, and 131, respectively. The number of protein spots that decreased by at least 50% following LPS exposure was 72, 151, 102, 98, 128, and 97, respectively. Although gel-to-gel regional variability in resolution was expected to account for individual spots not being well visualized on particular gels, only those spots that were matched to all 12 original gels were analyzed further. Overall, the number of spots matched to all 12 original gels was 125. The numbers of spots that were both matched to all 12 original gels and that increased by at least 50% in integrated intensity in the individual experiments following LPS exposure were 46, 13, 17, 27, 22, and 20, respectively. The numbers of spots that were matched to all 12 gels and that decreased by at least 50% were 6, 22, 17, 22, 34, and 28, respectively. The LPS-induced change in integrated intensity of the 125 spots that were matched to all 12 original gels was subjected to statistical analysis with a two-tailed Student's *t* test, and those spots with statistically significant ($p < 0.05$) regulation among the six experiments were identified by peptide mass fingerprinting (Table III).

Identification of LPS-regulated Proteins—Several proteins were consistently up-regulated on the pH 3.0–10.0 gels (Table III), including regulators of inflammation (annexin III) and signaling molecules (Rab-GDP dissociation inhibitor β). Several actin fragments were seen to be consistently up-regulated in the six experiments following LPS exposure (Table III). Of interest, the proteasome β chain was also consistently up-regulated. Down-regulated proteins included other signaling molecules, such as Rho GTPase activating protein 1.

On the pH 5.0–6.0 and 5.5–6.7 gels, several proteins were found to show increases of greater than 1.5-fold following LPS exposure (Tables IV and V), including cytoskeletal proteins, such as moesin, nonmuscle myosin heavy chain, and a putative phosphorylated form of nonmuscle myosin heavy chain, and signaling molecules, such as protein phosphatase 1 and PO_4-stathmin. The putative phosphorylated form of nonmuscle myosin heavy chain (*spot #1101*) was positioned 0.03 pH unit more acidic than the unmodified protein (*spot #1102*) (Fig. 1*D*) and was distinguished by a tryptic peptide (*m/z* 1366.74) not present in the unmodified protein, consistent with phosphorylation of serine 685. Serine 685 is predicted by NetPhos 2.0 Prediction Server (available at www.cbs.dtu.dk/services/NetPhos/(31)) to be a high probability phosphorylation residue and by ScanProsite (www.expasy.ch/tools/scnpsite.html) to be a substrate for protein kinase C. The tryptic phosphopeptide identified in PO_4-stathmin, extending from residues 15 to 27 (1468.7 Da), is consistent with phosphorylation of either serine 16, a known substrate for Ca^{2+}/calmodulin (CaM)-dependent kinases (32), or serine 25, a known substrate for p38δ and ERK (Fig. 2*A*)

TABLE I
Human neutrophil genes induced after 4 h of LPS exposure

Description	GenBank™ no.	Change-fold
Transcriptional regulation		
Pleiomorphic adenoma gene-like 2	D83784	16.8
NFKB2	S76638	12.3
NFKBIE	U91616	11.5
p65	L19067	8.4
BCL3	U05681	7.7
X-box binding protein 1	M31627	7.5
Metal-regulatory transcription factor 1	X78710	7.4
Ets-2	J04102	7.4
c-Rel	X75042	6.2
NFKB1	M58603	5.8
Basic leucine zipper transcription factor, ATF-like	U15460	4.7
IKB	M69043	3.8
MAX dimerization protein	L06895	3.6
DIF2	S81914	3.1
Cytokines and receptors		
MCP-1	M69203	78.7
MIP-1β	M72885	48.8
αHelix coiled-coil rod homolog	AF014958	20.8
IL-1β	X04500	17.6
GRO3 (beta)	M57731	17.3
TNF-α	X02910	14.5
MIP-3α	U64197	8.1
IL10RA	U00672	7.3
IL-6	Y00081	6.3
GROα	X54489	4
HM74	D10923	3.8
Immune response		
Orosomucoid	X02544	20.2
Complement component *C3*	K02765	12.8
Protease inhibitor 9	U71364	9.5
Complement component 3a receptor 1	U28488	6.1
Protease inhibitor 3	L10343	4.9
SLPI/antileukoprotease	X04470	4.7
ELANH2/elastase inhibitor	M93056	4.6
CD58	Y00636	3.8
Complement component *PFC*	M83652	3.5
Kinases		
CNK/FNK/PLK-like	U56998	16.2
Cot	D14497	11.9
Pim-2	U77735	9.5
LIMK2	D45906	4.3
Phosphatases		
PAC-1/DUSP2	L11329	11.8
DUSP5	U15932	5.3
PHA1	U73477	3.4
Signaling molecules		
TNFAIP1/A20	M59465	10
TRAF1	U19261	6.2
RanBP2	D42063	5.6
GNA15	M63904	5.2
PTAFR	D10202	3.9
Adhesion and cytoskeleton		
ICAM1	M24283	22.4
CEACAM1 (bilary glycoprotein)	X16354	6.3
LIMS1	U09284	6.1
SNL/actin bundling protein	U03057	5.9
Galectin-1/LGALS1	M57710	4.7
MEMD/ALCAM	U30999	4.2
CD44	HG2981—HT3125	3.9
TSG-6	M31165	3.7
Metabolic		
GTP cyclohydrolase I	U19523	13.5
NDUFV2/ubiquinone reductase	M22538	8.6
PSMA6/(proteosome iota)	X59417	8.4
UDP-galactose transporter (*SLC35A2*)	D84454	7.3
PLAU (urokinase)	X02419	6.4
KYNU/L-kynurenine hydrolase	U57721	5.5
AMPD3	D12775	5
P4HA1/prolyl 4-hydroxylase	M24486	4.7
γ *Glutamylcysteine synthetase*	L35546	4.5
ATP6D	J05682	4.2
ATP6S1	D16469	4

TABLE I—*continued*

Description	GenBank™ no.	Change-fold
Glycerol kinase	X68285	3.6
FACL1	L09229	3.5
AK3	X60673	3.3
Interferon-inducible		
ISG15	M13755	22.5
Mx1	M33882	19.4
IFI56	M24594	12.1
INDO	M34455	5.2
GBPI	M55542	4.3
PRKR	U50648	3.7
IFIT4	U52513	3.6
IFI54	M14660	3.5
IFI58	U34605	3.5
IFP35	U72882	3
Other		
Gos2	M72885	48.8
MIHC/IAP1	U37546	7.2
KIAA0105	D14661	5.1
KIAA0118	D42087	5
SNAP23	U55936	5
CASP5	U28015	4.8
KIAA0113	D30755	4.8
KIAA0255	D87444	4.7
Hepatoma-derived GF	D16431	4.7
PTGS2	D28235	4.6
CD48	M37766	4.3
UNC119 homolog	U40998	4.2
KIAA0151	D63485	3.9
Rab1b	XM035660	3.8
Annexin VII	J04543	3.7
KIAA0110	D14811	3.7
Adrenomedullin	D14874	3.7
AIM1	U83115	3.6
KIAA0250	D87437	3.2
P5-1	L06175	3.2
Scavenger receptor expressed by endothelial cells	D63483	3.2
VHL	L15409	3.1

(33). Assuming that no other multiply phosphorylated stathmin species had escaped detection, analysis of the integrated intensities of the PO_4-stathmin and stathmin spots indicates that the percentage of the PO_4 form of total cellular stathmin increased from 11% to 38% with LPS stimulation (Fig. 2*B*). This is similar to a previous report of an increase from <10% to 35–40% of the Ser^{25}-phosphorylated form in Jurkat cells stimulated with anti-CD3 (34).

Effect of SB203580 on LPS-stimulated Gene Expression—Gene expression analysis of PMNs stimulated with LPS indicated that the majority of genes induced by LPS were unaffected by prior treatment of PMN with SB203580. Of the 100 genes up-regulated by LPS, the up-regulation of 23 was inhibited by greater than 40% (Table VI). The majority of these genes affected by SB203580 were inhibited by less than 60%, whereas only six were inhibited by greater than 80%, all of which represent previously identified interferon-stimulated genes. Induction of cytokine genes by LPS, with the exception of *IL-6*, was generally unaffected by SB203580.

Effect of SB203580 on LPS-stimulated Protein Expression—Similar to the effect of SB203580 on LPS-stimulated gene expression, little effect of SB203580 was seen on expression levels for the majority of LPS-regulated proteins (Table VII). Two exceptions are annexin III and α-enolase, for which LPS-stimulated expression was attenuated in the presence of the p38 MAPK inhibitor.

Comparison of Microarray and Proteomics Results—Of the LPS-regulated proteins identified by peptide mass fingerprinting for which probes were present on the oligonucleotide microarray, poor concordance was found at the mRNA level (Table VIII). For 13 LPS-up-regulated proteins, 2 corresponding mRNA transcripts were up-regulated, 1 was down-regulated, 5 were unchanged, and 5 were not detected by the Affymetrix chip. For 5 down-regulated proteins, 3 corresponding transcripts were down-regulated, 1 was unchanged, and 1 was not detected. Varying patterns of LPS regulation emerge for those candidates detected at both the transcript and protein level. Proteasome β chain was up-regulated at both the transcript and protein levels (Table VIII), with no notable effect of SB203580 on expression at either level. Similarly, CAP1, RhoGAP1, and ficolin 1 were down-regulated at both the mRNA transcript and protein level (Table VIII), with no notable effect of SB203580. Annexin III was down-regulated at the transcript level and up-regulated at the protein level, with an inhibitory effect of SB203580 seen only at the protein level (Tables VII and VIII).

DISCUSSION

Interaction of bacterial LPS with the human PMN represents a model system for studying the activation and output of the innate immune system during infection and inflammation. A recent publication (35) describes the gene expression changes of a cultured monocytic cell line after infection by the Gram-positive bacterium *Listeria monocytogenes*. The cell wall components of Gram-positive bacteria, like Gram-negative-derived LPS (*i.e.* from *E. coli*), are known to signal through TLRs (36, 37). Importantly, many of the expression changes found in LPS-stimulated PMNs in the present study were also described in the bacteria-exposed monocytic cells, indicating that many of the gene expression changes seen in bacterial infection are likely mediated by TLRs (38, 39) and that the LPS model system accurately reflects exposure of immune cells to infec-

TABLE II
Human neutrophil genes repressed (>4-fold) after 4 h of LPS exposure

Description	GenBank™ no.	Change
		-fold
Kinases		
CAMK, II, gamma	U50360	−4
Diacylglycerol kinase, delta	D63479	−4.2
PRKCL2/PRK2 protein kinase C-like 2	U33052	−4.3
MAPKAPK3	U09578	−6.3
Protein kinase Ht31, cAMP-dependent	HG2167–HT2237	−8
CAMK II	L07044	−9.8
Transporters		
SLC25A5/solute carrier family 25, member 5	J02683	−4.2
SLC19A1; folate transporter	U17566	−4.4
SLC2A3; facilitated glucose transporter	M20681	−5
Metabolic		
Carbonic anhydrase IV	L10955	−4.4
RNase A family, k6	U64998	−4.5
Glycogen phosphorylase; liver	M14636	−4.6
Inositol polyphosphate-5-phosphatase	U57650	−4.6
Inositol 1,3,4-trisphosphate 5/6-kinase	U51336	−4.7
Transketolase	L12711	−4.8
Protein phosphatase 4, reg. subunit 1 (clone 23840)	U79267	−4.9
Cytidine deaminase	L27943	−5.4
MGAT1	M55621	−5.4
HMOX1	X06985	−5.4
MAN2A2	L28821	−5.8
Glycogenin (also represents U31525)	HG4334–HT4604	−5.9
Structural		
Fibrinogen-like protein (pT49 protein)	Z36531	−4.2
H2AFZ	M37583	−4.7
Paxillin	U14588	−4.9
Lamin B R	L25931	−5.9
Dynamin 2	L36983	−6.2
Actinin 1	M95178	−6.7
α-Tubulin	X01703	−10
Tubulin, α1, isoform 44	HG2259–HT2348	−15
Transcriptional regulators		
Lymphoblastic leukemia-derived sequence 1	M22638	−4.4
MAX-interacting protein 1	L07648	−4.5
Nuclear factor crythroid 2 isoform f	S77763	−6
Transducer of ERBB2, 1	D38305	−6.9
NFATC4	L41067	−7.8
ATF-2 (CRE-Bpa)	L05515	−9.6
Receptors		
Lymphotoxin β receptor	L04270	−4.4
Folate receptor 3 (gamma)	U08471	−5
	U11875	−5.3
Signaling		
Pix-α; cool-2 (KIAA0006)	D25304	−4.5
ARHB/RhoB	M12174	−4.5
TNFSF10; TRAIL	U37518	−6.6
Ca^{2+} binding		
ANXII	L19605	−4.3
S100A4	M80563	−4.8
ANX1	X05908	−4.8
Other		
Proteolipid protein 2	L09604	−4.9
Protein phosphatase 1, α catalytic subunit	HG1614–HT1614	−5
TIMP2	M32304	−5.1
KIAA0199	D83782	−5.2
Lipin 2 (KIAA0249)	D87436	−5.6
LRMP (Jaw1)	U10485	−5.8
CUGBP2	U69546	−6.9
Clone 23933	U79273	−7
PECAM1	L34657	−8
Delta sleep-inducing peptide	Z50781	−8.7
DiGeorge synd. critical region gene 2 (KIAA0163)	D79985	−9
SELPLG; CD162; selectin P ligand	U25956	−32

tion. Nevertheless, the reliance upon DNA microarrays alone affords insight only into the transcriptional response without corroboration at the protein level. In the present study, application of both DNA microarray and proteomics technology to our model system provides unique insight into both the cellular biology of the activated PMN and the responsiveness and reg-

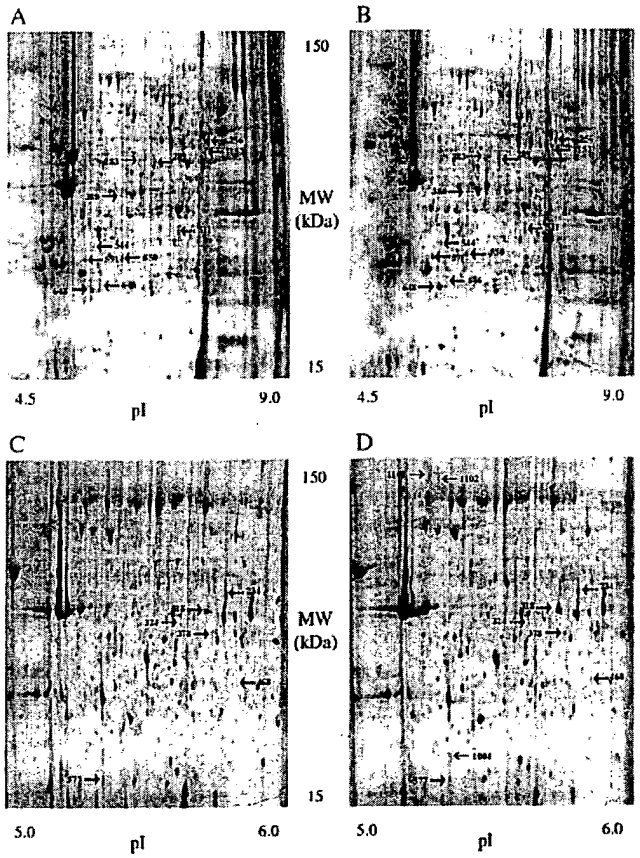


FIG. 1. **Two-dimensional PAGE of LPS-exposed human PMNs.** *A* and *B*, colloidal Coomassie Blue-stained pH 3.0–10.0, two-dimensional PAGE gels (*A*, control; *B*, LPS-exposed) with up-regulated (*solid arrows*) and down-regulated (*hatched arrows*) proteins indicated. These results are representative of six separate experiments. *C* and *D*, colloidal Coomassie Blue-stained pH 5.0–6.0, two-dimensional PAGE gels (*C*, control; *D*, LPS-exposed) with up-regulated (*solid arrows*), new (*solid arrow, open arrowhead*), and down-regulated (*hatched arrows*) proteins indicated. LPS-exposed PMNs from three blood donors were pooled.

ulation of its transcriptional and translational machinery. As will be discussed below, our study identifies, in particular, novel aspects of the LPS-stimulated PMN transcriptional regulation, activity in the innate immune response, signaling, cytoskeletal reorganization, and priming for granule release.

In the present study, the increase in NF-κB transcript abundance (Table I) detected by the microarrays corroborates the findings of other studies of PMNs and monocytes (40) and indicates a mechanism for the responsiveness and scope of the PMN transcriptional machinery following LPS exposure. NF-κB, recently described to be activated by LPS through the TLR/MyD88/interleukin-1 receptor-associated kinase pathway (1, 4), is the only transcriptional complex reported to be induced by LPS in the PMN. However, because the transcriptional NF-κB complex has been implicated in the regulation of only a portion of the genes induced by LPS in this study (data not shown), the importance of alternative transcriptional regulators in the PMN is clear. Of interest, several other known and putative transcriptional regulators with less well defined functions were also up-regulated in the present study, including *PLAGL2*, a putative zinc-finger protein, *XBP-1*, *MTF-1*, *Ets-2*, *B-ATF*, and *DIF-2*. On the other hand, LPS-down-regulated genes include *ATF-2* (a known target of p38), *NFATC4*, *TOB-1*, *NF-E2*, *MXI-1*, and *LYL-1*. Although the exact role of these gene products in regulating cell function is unknown, these data indicate that the range of transcriptional responses in the LPS-stimulated PMN is much broader than previously suggested and that the signaling capabilities of the PMN in the immune response are thereby likely extended in scope and specificity.

As expected from the literature, the genes for several cytokines and chemokines, including *IL-1β*, *IL-6*, and *MIP-1β*, were found to be up-regulated (Table I). On the other hand, the notable absence of up-regulated cytokines in the proteomics experiments reflects their removal in the post-LPS incubation wash performed prior to lysis for two-dimensional-PAGE. Up-regulation of these inflammatory mediators is well documented in PMNs exposed to LPS and in animal models of LPS-induced sepsis syndrome and acute respiratory distress syndrome, a PMN-mediated illness (41, 42). Several genes in this family were up-regulated that have not, to our knowledge, been described in LPS-stimulated cells, including *MCP-1*, *GRO3*, *IL-10RA*, and *HM74*, an orphan G protein-coupled receptor with homology to chemokine receptors. The down-regulation of *TNFSF10*, *lymphotoxin b receptor*, and *TNFAIP1* were also observed. The modulation of genes involved in cytokine signaling, including the adapter molecules *TRAF1* (LPS and TNF receptor signaling) and TNFAIP1 (TNF receptor signaling) and several kinases and phosphatases, may indicate a change in cytokine responsiveness after LPS treatment. Relevant in this regard from the proteomics data are: 1) the up-regulation of protein phosphatase 1, which has been shown to regulate PMN NADPH oxidase activation and translocation (43, 44) and to regulate LPS-induced NF-κB activation (45); 2) the down-regulation of Rho-GAP1, which has been shown to regulate NADPH oxidase activity in the PMN (46); and 3) the up-regulation of PO_4-stathmin (Table IV), a phosphoprotein postulated to function as a relayer and integrator of multiple signal transduction pathways (34). Several noncytokine, nonchemokine genes involved in the immune response were also up-regulated, including the complement pathway members *C3*, *C3AR1*, and *PFC*; the protease inhibitors *ELANH2* (elastase inhibitor), *SLPI*, *PI-3*, and *PI-9*; and the acute phase protein *orosomucoid*. LPS regulation of *C3AR1* and *orosomucoid* expression have not previously been reported. In the proteomics experiments, the down-regulation of ficolin-1 (Table III), a collectin-like cell surface protein reported to activate the complement system and to mediate adhesion and phagocytosis in monocytes but not previously reported in granulocytes (47), may represent negative modulation of the innate immune response. The finding that genes other than cytokines and chemokines are regulated by the PMN in response to LPS indicates that the PMN plays a more sophisticated role in host-defense and immunity than previously thought.

Treatment of the PMN with LPS lead to the induction of a set of genes associated with the anti-viral Type I interferons, IFNα/β. This induction occurs independently of the release of IFN or another unidentified soluble factor.[2] Furthermore, the set of genes expressed is smaller than that induced by IFNα/β, as described by Der *et al.* (12). This may be due to differences in the scope of the signaling systems activated by LPS and IFNα/β, or the time course of analysis of genes in the LPS-stimulated PMN. The implication that LPS treatment of PMN allows PMN to express anti-viral activity is currently being tested. Of interest was the finding that induction of interferon-stimulated genes was blocked by pretreatment of PMNs with SB203580. Work from our laboratory has indicated that signal transducers and activators of transcription activation does not occur in response to LPS in PMNs.[2] In addition, interferon-

[2] K. C. Malcolm and G. S. Worthen, manuscript in preparation.

TABLE III
Analysis of pH 3.0–10.0 two-dimensional PAGE gels

Mean change(-fold) in expression level among six PMN donors is reported. The change in expression for the proteins listed was statistically significant ($p < 0.05$) as measured by a two-tailed Student's t test.

Identification [spot no.]	Swiss-Prot no.	Estimated M_R/pI	Theoretical M_R/pI	Peptides matched/ submitted	Protein covered	Mean change
				%	%	-fold
Up-regulated						
Proteasome β chain [646]	P28070	27/5.7	29.2/5.72	9/12 (75%)	36%	1.51
Annexin III [550]	P12429	31/5.7	36.4/5.6	14/18 (78%)	42%	1.37
Actin *fragment* [544][a]	P02570	32/5.5	(41.7/5.29)	13/15 (87%)	(34%)	1.74
Actin *fragment* [591][a]	P02570	30/5.4	(41.7/5.29)	14/18 (78%)	(29%)	1.60
α-Enolase [380]	P06733	41/5.7	47.2/7.01	9/10 (90%)	24%	1.65
Rab-GDP dissociation inhibitor β [289]	P50395	50/6.1	50.7/6.11	10/11 (91%)	25%	1.24
Glutathione *S*-transferase P [648]	P09211	23/5.5	23.4/5.43	6/8 (75%)	41%	1.54
Pre-B-cell colony enhancing factor [1152]	P43490	53/7.0	55.5/6.69	12/16 (75%)	25%	1.29
Down-regulated						
Adenylyl cyclase-associated protein 1 [256]	Q01518	55/7.3	51.7/8.07	16/22 (73%)	34%	0.53
Rho-GAP1 [283]	Q07960	50/5.8	50.4/5.85	7/9 (78%)	22%	0.67
Ficolin 1 [511]	O00602	33/6.5	35/6.39	10/12 (83%)	25%	0.74

[a] The theoretical pI and M_R of native actin are indicated. Protein coverage indicates coverage of native actin.

TABLE IV
Analysis of pH 5.0–6.0 two-dimensional PAGE gels

Results are from pooled samples for control ($n=3$) and LPS-exposed ($n = 3$) PMNs from human donors. Expression of the reported proteins was altered >1.5-fold following LPS exposure in two repeat experiments. "New" designates proteins seen in the LPS gel in two repeat experiments but not detectable in the corresponding control gels.

Identification [spot no.]	Swiss-Prot no.	Estimated M_R/pI	Theoretical M_R/pI	Peptides matched/ submitted	Protein covered	Change
				%	%	-fold
Up-regulated						
Protein-tyrosine kinase 9-like [468]	Q9Y3F5[a]	34/5.81	39.5/6.37	10/14 (71%)	34%	1.8
Protein phosphatase 1, catalytic subunit, β isoform [378]	P37140	38/5.73	37.2/5.84	7/10 (70%)	22%	2.0
PO_4-stathmin [577]	P16949[b]	18/5.36	17.3/5.76	9/12 (75%)	42%	2.1[e]
Nonmuscle myosin heavy chain [1102]	189036[c]	145/5.32	145/5.23	20/21 (95%)	17%	New
Putative PO_4-nonmuscle myosin heavy chain [1101][d]	189036[b,c]	145/5.29	145/5.23	14/16 (87%)	13%	New
Leukocyte elastase inhibitor [318]	P30740	42/5.71	42.7/5.9	9/13 (69%)	22%	2.4
Grancalcin [1004]	P28676	24/5.36	24.0/5.02	7/10 (70%)	31%	New
Down-regulated						
Adenosylhomocysteinase [324]	P23526	48/5.82	47.7/6.04	7/9 (78%)	14%	0.4
PEST phosphatase interacting protein homolog [234][g]	4100162[f]	48/5.30	47.6/5.35	11/13 (85%)	30%	0.5

[a] TrEMBL accession number.
[b] Accession number and theoretical pI and M_R for the unmodified protein are indicated.
[c] *N*CBI accession number.
[d] See text for explanation.
[e] Among three experiments, the ratio of PO_4-stathmin expression increase, following LPS exposure in the presence of SB203580 divided by that in the absence of SB203580, was 0.93.
[f] Genpept accession number.
[g] This search was performed using average masses measured by linear mode MALDI-TOF MS.

TABLE V
Analysis of pH 5.5–6.7 two-dimensional PAGE gels

Results are from pooled samples for control ($n = 3$) and LPS-exposed ($n = 3$) PMNs from human donors. Expression of the reported proteins was altered >1.5-fold following LPS exposure in two repeat experiments.

Identification [spot no.]	Swiss-Prot no.	Estimated M_R/pI	Theoretical M_R/pI	Peptides matched/ submitted	Protein covered	Change
				%	%	-fold
Up-regulated						
Transaldolase [475]	P37837	38/5.95	37.5/6.36	13/17 (76%)	33%	2.5
Isocitrate dehydrogenase [431]	O75874	46/6.25	46.7/6.35	7/7 (100%)	13%	2.3
Moesin [201]	P26038	61/6.09	67.8/6.07	11/13 (85%)	17%	2.1
α-Enolase [459]	P06733	43/5.64	47.2/7.01	7/10 (70%)	17%	3.8
Down-regulated						
Calponin H2 [240]	Q99439	34/6.65	33.7/6.94	10/11 (90%)	27%	0.5

regulatory factor 3, a known regulator of interferon-stimulated gene transcription, is not a direct target of p38 kinase.[2] Therefore, gene expression analysis of LPS-stimulated PMNs has uncovered a previously uncharacterized signal transduction system that is sensitive to inhibition of p38 MAPK.

Knowledge of the genes down-regulated by LPS permits the development of further hypotheses addressing PMN function in the face of infection. Strikingly, several down-regulated genes and gene products are structural in nature (*e.g.* paxillin, actinin, calponin H2) (Tables II and V). A known consequence to the PMN of LPS exposure is decreased motility (48). Up-regulation of genes for adhesion molecules (*ICAM-1*, *CD44*, *AL-*

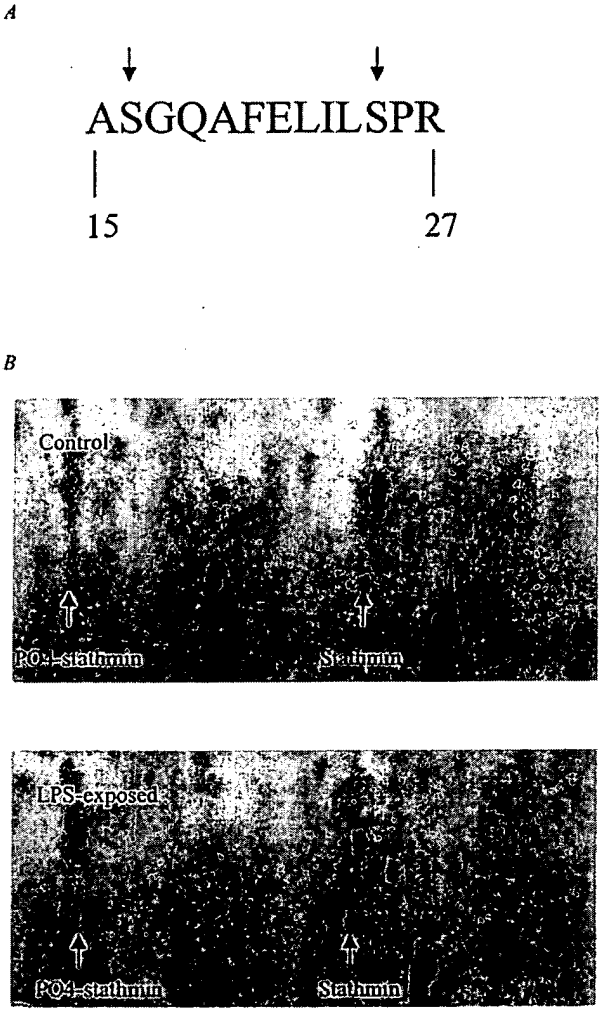


FIG. 2. *A*, the predicted sequence of the tryptic phosphopeptide in PO_4-stathmin (1468.72 Da). The peptide mass measured by MALDI-TOF MS and the predicted mass differed by 14 ppm. As indicated, two alternate phosphorylation sites are possible: serine 16 and serine 25. *B*, PO_4-stathmin and stathmin were identified on the control and LPS-exposed pH 5.0–6.0 gels. Consistent with phosphorylation, the PO_4-stathmin spot was distinguished by a peptide of mass 1468.72 Da (*i.e.* 80 Da greater than the peptide of 1388.72 Da seen in the stathmin spot). Assuming that no other multiply phosphorylated stathmin species have escaped detection, analysis of the integrated intensities of the PO_4-stathmin and stathmin spots indicates that the percentage of the PO_4 form of total cellular stathmin has increased from 11% to 38% with LPS stimulation. The decrease in integrated intensity for stathmin was equal in amount to the increase in PO_4-stathmin following LPS exposure.

CAM, and *TSG-6*), and down-regulation of genes for structural proteins, indicates a genetic basis for this observation. Down-regulation of two genes implicated in cytoskeletal regulation, *Pix-α* and *RhoB*, was also observed. The calcium-binding protein S100A4, down-regulated in LPS-treated PMNs (Table II), has been implicated in cell motility and metastasis (49). Decreased motility may be beneficial in sustaining the inflammatory response at sites of infection. In addition, LPS treatment results in an inhibition of apoptosis (50). Therefore, the longer residence time of the PMN at sites of infection is consistent with the long term genetically coded changes seen in these gene-profiling experiments and indicates that the changes in gene expression are functionally relevant to host defense and immunity.

By providing information on post-translational modification, the proteomics data may provide further insights into the cytoskeletal remodeling effects of LPS upon the PMN. We contend that the actin fragments identified (Table III) are unlikely to represent technical artifacts. Rather, their specificity (identical molecular weight/pI among different experiments), statistically significant up-regulation by LPS, as well as the use of a lysis buffer containing chaotropes and multiple protease inhibitors argue instead that these fragments are physiologic consequences of LPS exposure in the human PMN. More specifically, the up-regulation of these fragments following LPS exposure (Table III) suggests that LPS may activate an actin-cleaving enzyme, which, in turn, remodels the cytoskeleton. Intriguing in this vein, calpain has recently been reported to play an important role in cell migration and cytoskeletal organization of fibroblasts (51). The possibilities that LPS may induce calpain activation and that calpain activation may regulate cytoskeletal reorganization and motility are currently under investigation. An alternative possibility is that actin cleavage is a marker of neutrophil apoptosis (52).

Other LPS-regulated proteins may play important roles in cytoskeletal reorganization. The up-regulation of protein-tyrosine kinase 9-like (A6-related protein) may modulate LPS-

TABLE VI
Effect of SB203580 on LPS-stimulated gene expression

Genes are reported for which the SB203580/control expression ratio is ≤ 0.60.

Gene name	-fold change ratio (SB203580/control)	Change in absence of SB203580
		-fold
ISG15	0.09	22.5
HCR	0.38	20.8
Mx-1	0	19.4
IFI56	0	12.1
PI-9	0.57	9.5
Ets-2	0.59	7.4
IL-6	0.45	6.3
Rel	0.50	6.2
LIMS1	0.58	6.1
C3AR1	0.49	6.1
INDO	0.35	5.2
KIAA0105	0.41	5.1
SNAP23	0.58	5.0
SLPI	0.58	4.7
ELNAH2	0.49	4.6
HM-74	0.57	3.8
PKR	0	3.7
MAD	0.21	3.6
IFIT4	0.12	3.6
Glycerol kinase	0	3.6
IFI54	0	3.5
IFI58	0.39	3.5
IPF35	0.46	3.0

TABLE VII
Effect of SB203580 on LPS-stimulated protein expression

Protein name	-fold change ratio (SB203580/control)	Change in absence of SB203580
		-fold
Up-regulated		
Proteasome β chain	0.8	1.51
Annexin III	0.6	1.37
Actin *fragment* [544]	0.8	1.74
Actin *fragment* [591]	0.8	1.60
α-Enolase	0.6	1.65
Rab-GDP dissociation inhibitor β	1.1	1.24
Glutathione *S*-transferase P	1.2	1.54
Pre-B-cell colony enhancing factor	1.2	1.29
Down-regulated		
Adenylyl cyclase-associated protein 1	1.3	0.53
Rho-GAP1	0.8	0.67
Ficolin 1	1.0	0.74

TABLE VIII
LPS-regulated proteins for which a probe was present on the Affymetrix chip

A comparison of corresponding protein and mRNA transcript changes following LPS exposure is shown.

Protein	Protein change	mRNA change
		-fold
Up-regulated		
Proteasome β chain	1.5	1.9 ↑
Leukocyte elastase inhibitor	2.4	4.6 ↑
Rab-GDI β	1.24	NC[a]
Grancalcin	New	NC
Transaldolase	2.5	NC
Moesin	2.1	NC
Nonmuscle myosin heavy chain	New	NC
Glutathione *S*-transferase P	1.54	Absent
Pre-B cell enhancing factor	1.29	Absent
Isocitrate dehydrogenase	2.3	Absent
PO_4-stathmin	2.1	Absent (stathmin)
Protein phosphatase 1, β catalytic subunit	2	Absent
Annexin III	3.1	3.1 ↓
Down-regulated		
Adenylyl cyclase-associated protein 1	1.9	2.1 ↓
Rho-GAP 1	1.5	2.7 ↓
Ficolin 1	1.4	1.7 ↓
Adenosylhomocysteinase	2.5	Absent
Calponin H2	2	NC

[a] NC, no measureable change.

induced actin polymerization, because it bears a high degree of homology to twinfilin (A6), an actin monomer-binding protein that localizes to sites of rapid filament assembly in cells and is believed to regulate actin filament turnover (53). In turn, LPS-induced down-regulation of Rho-GTPase activating protein 1 (Table III) may regulate twinfilin (and protein-tyrosine kinase 9-like) activity, because twinfilin has been shown to colocalize with Rac1 and Cdc42 and to be regulated by active Rac1 in NIH 3T3 cells (53). Activation of Rho proteins may be facilitated by LPS up-regulation of moesin (Table V), because moesin reportedly induces the dissociation of Rho from GDI (54). Rac1 may, in turn, promote activation of the actin filament-nucleating Arp2/3 complex through interactions with WASP (Wiskott-Aldrich syndrome protein) family proteins (55) and, interestingly, is postulated to regulate the dynamics of both the actin and microtubule cytoskeletons via phosphorylation of stathmin (Table IV) (56). Calponin H2 is an actin-binding protein not previously reported in PMNs that is postulated to play a role in cytoskeletal organization (57). Its down-regulation by LPS (Table V) likely modulates LPS-induced cytoskeletal reorganization. The up-regulation of nonmuscle myosin heavy chain and a putative phosphorylated form of myosin heavy chain (putative protein kinase C substrate by prediction rules) in the LPS-exposed PMN (Table IV) is of uncertain significance; myosin has been implicated in multiple functions in the PMN, including locomotion, fluid pinocytosis, and phagocytosis (58). Of interest, however, S100A4 (down-regulated, Table II) has been reported to regulate cytoskeletal dynamics by inhibiting protein kinase C-mediated phosphorylation of nonmuscle myosin heavy chain (59).

LPS induction of stathmin phosphorylation (Table IV and Fig. 2) may represent another mechanism by which the cytoskeleton is remodeled. Stathmin is a phosphoprotein reportedly involved in both signal transduction and in regulation of the microtubulin filament network; furthermore, phosphorylation of stathmin has been reported to modulate its tubulin-binding avidity (60). Inferences can be made about both the phosphorylation site on PO_4-stathmin and the responsible kinase induced by LPS. Four phosphorylation sites in stathmin have been well described: Ser^{16}, Ser^{25}, Ser^{38}, and Ser^{63} (32, 33). Ser^{16} has been reported as a substrate for Ca^{2+}/calmodulin (CaM)-dependent kinases (32), and Ser^{25} as primarily a substrate for p38 and ERK (33), with $p34^{cdc2}$ also active but bearing a 5-fold preference for Ser^{38} (34). As stated above, the phosphopeptide identified in PO_4-stathmin, extending from residues 15 to 27 (1468.7 Da), is consistent with phosphorylation of either Ser^{16} or Ser^{25} (Fig. 2). Although both p38δ and p38α MAPK isoforms are expressed in the human PMN, LPS has been shown to selectively activate the p38α isoform in human PMNs (9). The p38α isoform, however, has been shown to be relatively inactive at Ser^{25}; in fact, p38δ is ~100-fold more active at Ser^{25}, and selective p38α inhibitors do not inhibit the stress-activated phosphorylation of stathmin in 293 cells (33). Further support for the lack of involvement of p38 signaling in phosphorylation of stathmin in our system is the apparent lack of effect of SB203580 (a selective p38α and p38β inhibitor) on LPS-induced expression of PO_4-stathmin (Table IV). Because $p34^{cdc2}$ is relatively inactive at Ser^{25} (34), we conclude that the phosphorylation site is likely to be Ser^{16}, a reported substrate of CaM-dependent kinase. Although CaM kinases have previously been implicated in gene activation in LPS-exposed myelomonocytic HD11 cells (61), stathmin signaling has not, to our knowledge, been previously reported in either PMNs or lipopolysaccharide signal transduction.

Cytoskeletal reorganization, a well-described regulator of granule release (62), may underlie LPS-induced priming for PMN granule release, but several LPS-regulated proteins may provide more specific clues. LPS exposure led to increased levels of grancalcin, a calcium-binding protein previously detected in PMNs and shown to translocate to granules and plasma membrane in the presence of physiologic concentrations of calcium (63). Similarly, annexin III, a calcium-binding protein highly expressed in PMN granule membranes and implicated in calcium-mediated secretion (64) and in granule fusion (65), was also found to be up-regulated. Exocytosis of granule contents may also be facilitated by LPS up-regulation of Rab-GDP dissociation inhibitor (Table III), which has been proposed to recycle Rab after vesicle fusion by extracting it from the membrane and loading it onto newly formed transport intermediates (66).

Parallel use of DNA microarrays and proteomics affords a powerful strategy for comparison of corresponding mRNA transcripts and proteins, thereby affording new insight into the mechanisms by which the cell regulates its signaling responses to the external environment. Of interest, a poor correlation was found between corresponding transcripts and proteins (Table VIII), as reported in other systems (17, 18). The finding in some cases of unchanged transcript abundance in the face of regulated protein levels indicates post-transcriptional modulation following LPS exposure. The finding of undetected transcripts in the face of regulated levels of the corresponding proteins may indicate previous transcription of these genes in an earlier state of the myeloid maturation of the PMN, producing stable protein species that have undergone post-translational alteration following LPS exposure. The use of SB203580, a p38 inhibitor, adds further insights into the mechanisms of LPS regulation. At the level of mRNA expression, SB203580 inhibited 23% of LPS-stimulated genes by ≥40% and 11% of genes by ≥60%; therefore, p38 plays a specific role in gene regulation in the PMN. In particular, proteasome β chain was up-regulated at both the mRNA transcript and protein level (Table VIII), with no notable effect of SB203580 on expression at either level, consistent with a non-p38-mediated pathway of primary transcriptional up-regulation induced by LPS. Similarly, CAP1, Rho-GAP1, and ficolin 1 were down-regulated at both the mRNA transcript and protein level (Table VIII), with

no notable effect of SB203580, consistent with a non-p38-mediated pathway of primary transcriptional down-regulation. Interestingly, annexin III was down-regulated at the transcript level and up-regulated at the protein level, with an inhibitory effect of SB203580 seen only at the protein level (Table VII), consistent with a p38-mediated post-transcriptional up-regulation induced by LPS.

Limitations of the present study should be noted. Gene expression analysis by cDNA microarrays does not distinguish between transcriptional regulation and mRNA stabilization; similarly, two-dimensional PAGE proteomics by itself does not distinguish among transcriptional, translational, or post-translational regulation of protein abundance. Transcript detection by microarray technology is limited to the probes included; protein identification by two-dimensional PAGE proteomics is limited to well-resolved regions of the gel, may perform less well with hydrophobic and high molecular weight proteins, and tends to select for more abundant protein species (30). Harvesting of the LPS-incubated PMNs at 4 h may have prevented detection of earlier, transient changes and may have thereby introduced artifactual transcript-protein discordance. Furthermore, the post-LPS incubation, pre-two-dimensional PAGE cell washes would be expected to remove secreted proteins from further analysis, with uncertain effects on detected protein abundance depending on such factors as the degree of *de novo* synthesis and extent of degranulation/exocytosis. Because protein binding of Coomassie Blue has a limited dynamic range and is typically not linear throughout the range of detection, image analysis of Coomassie Blue-stained protein spots should be considered semi-quantitative. For some protein spots, the apparent magnitude of regulation by LPS may have been blunted by the spot approaching staining saturation in the control gel. By limiting our analysis to those protein spots common to all twelve pH 3.0–10.0 two-dimensional gels, we likely excluded some LPS-regulated proteins that happened to be either poorly resolved on a subset of the gels or unmatched by the image analysis software. By further limiting the analysis to those matched spots on the pH 3.0–10.0 gels for which a two-tailed t test demonstrated $p < 0.05$, the list of regulated proteins was likely also limited by statistical power. In addition to those regulated proteins listed in Table III, three others were up-regulated and three down-regulated with $p < 0.09$ (data not shown).

Limiting our reported results to those changes that met statistical significance among the donors carries further important implications. We have encountered a two order of magnitude range of response in unselected donor LPS-induced PMN functions, such as TNF-α and superoxide anion release (data not shown). The sources of this physiologic heterogeneity remain uncertain but may possibly include such factors as natural mutations of the LPS receptor component, TLR4 (67). By selecting for LPS effects common to all donors, we may not have characterized the range of genomic and proteomic heterogeneity present in the population and thereby may have focused on only a narrow portion of a broader biological response to LPS. We contend that this reductionist approach is valid because it would be expected to enrich for biologically integral responses of the PMN to LPS. Nevertheless, correlation of genomic and proteomic profiles with functional phenotypes of the PMN may bear important diagnostic and therapeutic implications and will be pursued in future studies.

Widespread regulation of numerous noncytokine/chemokine genes and proteins in the LPS-stimulated human PMN is a novel finding. These data indicate that, despite a narrow scope of gene expression in the nonstimulated state, the terminally differentiated, short-lived PMN likely plays a role in the innate immune response that is far more sophisticated and dynamic than the simple release of preformed inflammatory mediators. Although gene expression appears to be an important mechanism by which PMNs respond acutely to infection, mRNA transcript/protein concordance is limited, and post-transcriptional (and post-translational) modifications also play an important role. The alteration of multiple transcriptional regulators, G-protein regulators, PO_4-stathmin, and protein phosphatase 1 indicates that one of the responses to LPS exposure is to modify subsequent signaling events by bacterial components or by other cytokines and chemokines. Finally, the finding that p38 MAPK mediates LPS regulation of a limited subset of transcripts and proteins underlines the continuing need to define signal transduction cascades in the neutrophil.

Acknowledgments—We thank the members of the Affymetrix core laboratory, University of Colorado Health Sciences Center, as well as Benjamin Perryman, Steve Helmke, and Jennifer Lynch of the Cardiology Division, University of Colorado Health Sciences Center for assistance with two-dimensional PAGE.

REFERENCES

1. Bowie, A., and O'Neill, L. A. (2000) *J. Leukocyte Biol.* **67,** 508–514
2. Chow, J. C., Young, D. W., Golenbock, D. T., Christ, W. J., and Gusovsky, F. (1999) *J. Biol. Chem.* **274,** 10689–10692
3. Kirschning, C. J., Wesche, H., Merrill Ayres, T., and Rothe, M. (1998) *J. Exp. Med.* **188,** 2091–2097
4. Muzio, M., Polentarutti, N., Bosisio, D., Prahladan, M. K., and Mantovani, A. (2000) *J. Leukocyte Biol.* **67,** 450–456
5. Yang, R. B., Mark, M. R., Gray, A., Huang, A., Xie, M. H., Zhang, M., Goddard, A., Wood, W. I., Gurney, A. L., and Godowski, P. J. (1998) *Nature* **395,** 284–288
6. Parsons, P. E., Worthen, G. S., Moore, E. E., Tate, R. M., and Henson, P. M. (1989) *Am. Rev. Respir. Dis.* **140,** 294–301
7. Nick, J. A., Avdi, N. J., Young, S. K., McDonald, P. P., Billstrom, M. A., Henson, P. M., Johnson, G. L., and Worthen, G. S. (1999) *Chest* **116,** 54S–55S
8. Nick, J. A., Young, S. K., Brown, K. K., Avdi, N. J., Arndt, P. G., Suratt, B. T., Janes, M. S., Henson, P. M., and Worthen, G. S. (2000) *J. Immunol.* **164,** 2151–2159
9. Nick, J. A., Avdi, N. J., Young, S. K., Lehman, L. A., McDonald, P. P., Frasch, S. C., Billstrom, M. A., Henson, P. M., Johnson, G. L., and Worthen, G. S. (1999) *J. Clin. Invest.* **103,** 851–858
10. Cassatella, M. A. (1995) *Immunol. Today* **16,** 21–26
11. McDonald, P. P., Bald, A., and Cassatella, M. A. (1997) *Blood* **89,** 3421–3433
12. Der, S. D., Zhou, A., Williams, B. R., and Silverman, R. H. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 15623–15628
13. Karpf, A. R., Peterson, P. W., Rawlins, J. T., Dalley, B. K., Yang, Q., Albertsen, H., and Jones, D. A. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 14007–14012
14. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999) *Science* **283,** 83–87
15. Han, J., Jiang, Y., Li, Z., Kravchenko, V. V., and Ulevitch, R. J. (1997) *Nature* **386,** 296–299
16. Lee, J. C., Laydon, J. T., McDonnell, P. C., Gallagher, T. F., Kumar, S., Green, D., McNulty, D., Blumenthal, M. J., Heys, J. R., Landvatter, S. W. *et al.* (1994) *Nature* **372,** 739–746
17. Anderson, L., and Seilhamer, J. (1997) *Electrophoresis* **18,** 533–537
18. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) *Mol. Cell. Biol.* **19,** 1720–1730
19. Lewis, T. S., Hunt, J. B., Aveline, L. D., Jonscher, K. R., Louie, D. F., Yeh, J. M., Nahreini, T. S., Resing, K. A., and Ahn, N. G. (2000) *Mol. Cell* **6,** 1343–1354
20. Soskic, V., Gorlach, M., Poznanovic, S., Boehmer, F. D., and Godovac-Zimmermann, J. (1999) *Biochemistry* **38,** 1757–1764
21. Triantafilou, K., Triantafilou, M., and Dedrick, R. L. (2001) *Nat. Immunol.* **2,** 338–345
22. Nick, J. A., Avdi, N. J., Young, S. K., Knall, C., Gerwins, P., Johnson, G. L., and Worthen, G. S. (1997) *J. Clin. Invest.* **99,** 975–986
23. Rabilloud, T., Valette, C., and Lawrence, J. J. (1994) *Electrophoresis* **15,** 1552–1558
24. Bradford, M. M. (1976) *Anal. Biochem.* **72,** 248–254
25. Neuhoff, V., Arold, N., Taube, D., and Ehrhardt, W. (1988) *Electrophoresis* **9,** 255–262
26. Arnott, D., O'Connell, K. L., King, K. L., and Stults, J. T. (1998) *Anal. Biochem.* **258,** 1–18
27. Hellman, U., Wernstedt, C., Gonez, J., and Heldin, C. H. (1995) *Anal. Biochem.* **224,** 451–455
28. Yan, J. X., Sanchez, J. C., Rouge, V., Williams, K. L., and Hochstrasser, D. F. (1999) *Electrophoresis* **20,** 723–726
29. Mardian, J. K., and Isenberg, I. (1978) *Anal. Biochem.* **91,** 1–12
30. Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., and Sanchez, J. C. (2000) *Electrophoresis* **21,** 1104–1115
31. Blom, N., Gammeltoft, S., and Brunak, S. (1999) *J. Mol. Biol.* **294,** 1351–1362
32. le Gouvello, S., Manceau, V., and Sobel, A. (1998) *J. Immunol.* **161,** 1113–1122
33. Parker, C. G., Hunt, J., Diener, K., McGinley, M., Soriano, B., Keesler, G. A., Bray, J., Yao, Z., Wang, X. S., Kohno, T., and Lichenstein, H. S. (1998)

Biochem. Biophys. Res. Commun. **249,** 791–796
34. Marklund, U., Brattsand, G., Shingler, V., and Gullberg, M. (1993) *J. Biol. Chem.* **268,** 15039–15047
35. Cohen, P., Bouaboula, M., Bellis, M., Baron, V., Jbilo, O., Poinot-Chazel, C., Galiegue, S., Hadibi, E. H., and Casellas, P. (2000) *J. Biol. Chem.* **275,** 11181–11190
36. Yoshimura, A., Lien, E., Ingalls, R. R., Tuomanen, E., Dziarski, R., and Golenbock, D. (1999) *J. Immunol.* **163,** 1–5
37. Lien, E., Sellati, T. J., Yoshimura, A., Flo, T. H., Rawadi, G., Finberg, R. W., Carroll, J. D., Espevik, T., Ingalls, R. R., Radolf, J. D., and Golenbock, D. T. (1999) *J. Biol. Chem.* **274,** 33419–33425
38. Underhill, D. M., Ozinsky, A., Hajjar, A. M., Stevens, A., Wilson, C. B., Bassetti, M., and Aderem, A. (1999) *Nature* **401,** 811–815
39. Brightbill, H. D., Libraty, D. H., Krutzik, S. R., Yang, R. B., Belisle, J. T., Bleharski, J. R., Maitland, M., Norgard, M. V., Plevy, S. E., Smale, S. T., Brennan, P. J., Bloom, B. R., Godowski, P. J., and Modlin, R. L. (1999) *Science* **285,** 732–736
40. de Wit, H., Dokter, W. H., Koopmans, S. B., Lummen, C., van der Leij, M., Smit, J. W., and Vellenga, E. (1998) *Leukemia* **12,** 363–370
41. Johnston, C. J., Finkelstein, J. N., Gelein, R., and Oberdorster, G. (1998) *Toxicol. Sci.* **46,** 300–307
42. Ulich, T. R., Watson, L. R., Yin, S. M., Guo, K. Z., Wang, P., Thang, H., and del Castillo, J. (1991) *Am. J. Pathol.* **138,** 1485–1496
43. Karlsson, A., Nixon, J. B., and McPhail, L. C. (2000) *J. Leukocyte Biol.* **67,** 396–404
44. Dorseuil, O., Quinn, M. T., and Bokoch, G. M. (1995) *J. Leukocyte Biol.* **58,** 108–113
45. Pahan, K., Sheikh, F. G., Namboodiri, A. M., and Singh, I. (1998) *J. Biol. Chem.* **273,** 12219–12226
46. Geiszt, M., Dagher, M. C., Molnar, G., Havasi, A., Faure, J., Paclet, M. H., Morel, F., and Ligeti, E. (2001) *Biochem. J.* **355,** 851–858
47. Teh, C., Le, Y., Lee, S. H., and Lu, J. (2000) *Immunology* **101,** 225–232
48. Wagner, J. G., and Roth, R. A. (1999) *J. Leukocyte Biol.* **66,** 10–24
49. Barraclough, R. (1998) *Biochim. Biophys. Acta* **1448,** 190–199
50. Lee, A., Whyte, M. K., and Haslett, C. (1993) *J. Leukocyte Biol.* **54,** 283–288
51. Dourdin, N., Bhatt, A. K., Dutt, P., Greer, P. A., Arthur, J. S., Elce, J. S., and Huttenlocher, A. (2001) *J. Biol. Chem.* **276,** 48382–48388
52. Brown, S. B., Bailey, K., and Savill, J. (1997) *Biochem. J.* **323,** 233–237
53. Vartiainen, M., Ojala, P. J., Auvinen, P., Peranen, J., and Lappalainen, P. (2000) *Mol. Cell. Biol.* **20,** 1772–1783
54. Takahashi, K., Sasaki, T., Mammoto, A., Takaishi, K., Kameyama, T., Tsukita, S., and Takai, Y. (1997) *J. Biol. Chem.* **272,** 23371–23375
55. Machesky, L. M., and Insall, R. H. (1998) *Curr. Biol.* **8,** 1347–1356
56. Daub, H., Gevaert, K., Vandekerckhove, J., Sobel, A., and Hall, A. (2001) *J. Biol. Chem.* **276,** 1677–1680
57. Masuda, H., Tanaka, K., Takagi, M., Ohgami, K., Sakamaki, T., Shibata, N., and Takahashi, K. (1996) *J Biochem. (Tokyo)* **120,** 415–424
58. Valerius, N. H., Stendahl, O. I., Hartwig, J. H., and Stossel, T. P. (1982) *Adv. Exp. Med. Biol.* **141,** 19–28
59. Kriajevska, M., Tarabykina, S., Bronstein, I., Maitland, N., Lomonosov, M., Hansen, K., Georgiev, G., and Lukanidin, E. (1998) *J. Biol. Chem.* **273,** 9852–9856
60. Steinmetz, M. O., Jahnke, W., Towbin, H., Garcia-Echeverria, C., Voshol, H., Muller, D., and van Oostrum, J. (2001) *EMBO Rep.* **2,** 505–510
61. Regenhard, P., Goethe, R., and Phi-van, L. (2001) *J. Leukocyte Biol.* **69,** 651–658
62. Valentijn, K., Valentijn, J. A., and Jamieson, J. D. (1999) *Biochem. Biophys. Res. Commun.* **266,** 652–661
63. Teahan, C. G., Totty, N. F., and Segal, A. W. (1992) *Biochem. J.* **286,** 549–554
64. Rosales, J. L., and Ernst, J. D. (1997) *J. Immunol.* **159,** 6195–6202
65. Le Cabec, V., and Maridonneau-Parini, I. (1994) *Biochem. J.* **303,** 481–487
66. Gilbert, P. M., and Burd, C. G. (2001) *J. Biol. Chem.* **276,** 8014–8020
67. Arbour, N. C., Lorenz, E., Schutte, B. C., Zabner, J., Kline, J. N., Jones, M., Frees, K., Watt, J. L., and Schwartz, D. A. (2000) *Nat. Genet.* **25,** 187–191

Genomic and proteomic analysis of the myeloid differentiation program

Zheng Lian, Le Wang, Shigeru Yamaga, Wesley Bonds, Y. Beazer-Barclay, Yuval Kluger, Mark Gerstein, Peter E. Newburger, Nancy Berliner, and Sherman M. Weissman

Although the mature neutrophil is one of the better characterized mammalian cell types, the mechanisms of myeloid differentiation are incompletely understood at the molecular level. A mouse promyelocytic cell line (MPRO), derived from murine bone marrow cells and arrested developmentally by a dominant-negative retinoic acid receptor, morphologically differentiates to mature neutrophils in the presence of 10 μM retinoic acid. An extensive catalog was prepared of the gene expression changes that occur during morphologic maturation. To do this, 3′-end differential display, oligonucleotide chip array hybridization, and 2-dimensional protein electrophoresis were used. A large number of genes whose mRNA levels are modulated during differentiation of MPRO cells were identified. The results suggest the involvement of several transcription regulatory factors not previously implicated in this process, but they also emphasize the importance of events other than the production of new transcription factors. Furthermore, gene expression patterns were compared at the level of mRNA and protein, and the correlation between 2 parameters was studied. (Blood. 2001;98:513-524)

Introduction

Studies of normal myeloid maturation from many laboratories have identified genes that may play critical roles in myeloid differentiation.[1-4] Current studies suggest that these events are dependent on a cascade of molecular changes that involve complex modulation of mRNA transcription. Furthermore, studies of acute leukemia have suggested that the disease arises from the accumulation of myeloid precursors arrested at early stages of differentiation and associated, in many cases, with chromosomal rearrangements that alter the structure of specific transcription factors.[5] Nevertheless, the molecular events underlying the production of mature myeloid cells are not well understood and appear to use interacting pathways and networks, the elucidation of which requires an extensive description of the molecular components available to the myeloid cell.

An extensive body of information is accumulating with respect to gene expression profiles of mammalian cells. However, much of the information available in public databases has been accumulated by the use of techniques such as single oligonucleotide chips or cDNA arrays that measure fewer than 6000 of potentially 30 000 to 120 000 transcripts. The more limited range of analyses reported by the serial analysis of gene expression (SAGE)[6,7] technique accurately estimates changes in levels of the more abundant mRNAs but requires extensive redundant analyses to measure changes in the patterns of expression of scarce mRNAs. We have used a modified polymerase chain reaction (PCR)–based cDNA differential display (DD) method in which single restriction fragments derived from the 3′ end of cDNAs are separated on a sequencing gel.[8,9] Bands from the gel can be identified initially by sequencing, but then comparison of patterns from different samples can be made without further sequencing. This sensitive and reproducible method detects, in principle, most cDNAs regardless of whether they are represented in existing databases.

Systematic analysis of the function of genes can also be performed at the protein level. This approach has the advantage of being closest to function, because proteins perform most of the reactions necessary for the cell. The most common method of proteome analysis is the combination of 2-dimensional gel electrophoresis (2DE) to separate and visualize protein and mass spectrometry (MS) for protein identification.[10] Several such analyses of yeast and of normal or malignant mammalian cells have been performed. To date, however, there have been few studies in which both mRNA and protein have been compared by applying analyses to the same samples. The studies of Anderson[11] and Gygi[12] showed that there is not a good correlation between mRNA and protein levels, in yeast or human liver cells. However, other analyses disagree with this conclusion (Greenbaum et al, manuscript submitted, and Futcher et al[14]). Furthermore, global correlations between changes in mRNA and protein levels have not been examined during the execution of any developmental program.

The MPRO cell line was derived by transduction of a dominant-negative retinoic acid receptor construct into normal mouse bone marrow cells. It is a granulocyte-macrophage colony-stimulating factor (GM-CSF)–dependent line arrested at a promyelocytic stage of development.[15,16] After treatment with all-*trans* retinoic acid (ATRA) most of the cells acquire the morphology of mature

From the Department of Genetics, Boyer Center for Molecular Medicine, the Section of Hematology, Department of Internal Medicine, and the Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, CT; the Department of Pediatrics, University of Massachusetts Medical School, Worcester, MA; and Gene Logic, Gaithersburg, MD.

Submitted December 4, 2001; accepted March 28, 2001.

Supported by grants from the National Institutes of Health (NIH) (CA42556) and Gene Logic (A143558, DK54369, and HL63357). Z.L. is supported by NIH grant HL 63357. P.E.N. is supported by NIH grant DK 54369, grants from the Arthritis Foundation and the Charles H. Hood Foundation, and the Pierce Family Cancer Research Fund. M.G. is supported by the Keck Foundation and by NIH grant GM54160-04.

L.W. and S.Y. contributed equally to this research.

Reprints: Sherman M. Weissman, Department of Genetics, Boyer Center for Molecular Medicine, Yale University School of Medicine, Rm 336, 295 Congress Ave, New Haven, CT 06536-0812; e-mail: sherman.weissman@yale.edu.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 U.S.C. section 1734.

neutrophils and begin to produce neutrophil lactoferrin and gelatinase, 2 proteins characteristic of neutrophil secondary granules.[17] As such, it offers a valuable model for studying neutrophil differentiation in vitro.

We now report the analysis of mRNA expression changes during the process of MPRO cell maturation to neutrophils and compare the results with a limited analysis of cellular protein composition. mRNA expression changes were studied by combining the use of oligonucleotide arrays and DD. A database (dbMC) with comprehensive genomic information for myeloid differentiation program was constructed (accessible at http://www.bioinfo.mbb.yale.edu/expression/neutrophil). We have grouped the changes in mRNA levels of a large number of genes into 6 patterns, with implications for the genetic program of myeloid differentiation.

We also compared 2-dimensional high-resolution gel electrophoretograms from control cells and cells differentiated for 72 hours in the presence of ATRA. Fifty protein spots whose relative intensity changed prominently during differentiation were examined by mass spectrometry. The results suggest a poor correlation between mRNA expression and protein abundance, indicating that it may be difficult to extrapolate directly from individual mRNA changes to corresponding ones in protein levels (as estimated from 2DE).

Materials and methods

Cell lines

MPRO cells and HM-5 cells provided by Dr Schickwann Tsai (Fred Hutchinson Cancer Research Center, Seattle, WA)[15] were used throughout the study. The cells proliferated continuously as a GM-CSF–dependent cell line at 37°C in Iscoves modified Dulbecco medium (Gibco BRL, Grand Island, NY) supplemented with 5% to 10% fetal calf serum (Gibco BRL) and 10% HM-5–conditioned medium as a source of GM-CSF. Morphologic differentiation of the blocked MPRO promyelocytes was induced by treatment with 10 μM ATRA (Sigma, St Louis, MO). Controls were cultured in the absence of ATRA but with the same volume of vehicle (ethanol).

RNA isolation and differential display

After exposure to 10 μM ATRA for 0, 24, 48, or 72 hours, total cellular RNA was isolated from MPRO cells using TRIzol reagent (Life Technologies, Gaithersburg, MD). cDNA was then synthesized using a T-7 Sal-Oligo d(T) 32 primer as described previously.[8,18] The double-stranded cDNA was digested with 1 of 9 different restriction enzymes (*Apa*I, *Bgl*II, *Bam*HI, *Eag*I, *Eco*RI, *Hind*III, *Xba*I, *Kpn*I, and *Sph*I) and ligated to Y-shaped adaptors with a complementary overhang. DNA fragments were then amplified by PCR as described previously.[8,18] PCR products were separated on a sequencing gel of 6% polyacrylamide with 7 M urea. The gel was dried and exposed to x-ray film. Genes from differential display gels, whose maximum intensity changes equaled 2+ on a scale of 1+ to 8+, were recorded as significantly changed.[19] Individual DNA bands were recovered from the gels, amplified by PCR, and sequenced.

Oligonucleotide chip analysis of RNA samples

Ten micrograms total RNA from each sample (0, 24, 48, or 72 hours) was used to prepare cDNA. This cDNA was transcribed with T7 RNA polymerase to prepare a fluorescently labeled probe.[20,21] Each sample was hybridized to mouse array chip (Mu11K Array; Affymetrix, Santa Clara, CA) containing oligonucleotide probe sets corresponding to approximately 7000 known genes or ESTs represented by UniGene clusters.[22] cDNAs were considered present if their probe set results were rated as such by the GeneChip software (Affymetrix) and if the average difference (AD) between perfect match and mismatch probe pairs was not less 100 U. If a gene was represented by more than one array probe set, the average of all probe sets for the gene was taken. Genes with AD values between 100 and 200 were considered unchanged because of their low expression levels. Those genes with AD values equal to or more than 200 U at one time point were further studied by rescaling, threshold, and normalization methods described in the MIT Center for Genome Research Web site.[13] A value of 20 was assigned to any gene with an AD below 20 at some time point.

Bioinformatics and database development

All the sequences or gene fragments were searched using Blast against GenBank and TIGR gene indices. A database of genes or ESTs whose expression levels changed during myeloid differentiation was constructed containing information for each band or gene. This included GenBank matches, Locus Link or Unigene clusters, expression patterns, tissue distribution, synonym(s) protein name, gene name(s), notations of possible functions, poly A signal and sequence quality, and hyperlinks to the database searches, sequence trace files, and related references. All gene data were then gathered into a cluster file. Supplementary information is available at http://bioinfo.mbb.yale.edu/expression/neutrophil.

Classification and analysis of DNA fragments

Sequences from differential display analyses were classified as representing known genes, ESTs, genomic sequences, or novel genes as described.[19,23] Known genes from both differential display and arrays were clustered into 27 functional categories and searched against SWISS-PROT (http://www.expasy.cbr.nrc.ca/cgi-bin/sprot-search-ful) or PIR (http://www.pir.georgetown.edu/). Information such as function, subcellular location, family and superfamily classification, map position, similarity, synonym(s) protein name, gene name(s), and so on was recorded in a variety of databases.

Northern blot analysis

Thirty micrograms total cellular RNA per lane from time-course MPRO cells were loaded onto 1.2% formaldehyde-agarose gels, then transferred to Hybond-N+ membranes (Amersham Pharmacia Biotech, Uppsala, Sweden). After standard prehybridization, membranes were hybridized overnight at 65°C with radiolabeled cDNA probes (ordered from Research Genetics according to their dbEST Image ID). Membranes were washed at a final stringency of 60°C in 0.1 × SSC.

Immobilized pH gradient 2-dimensional gel electrophoresis and mass spectrometry

Induced MPRO cells collected at 0 and 72 hours were lysed with lysis buffer (540 mg urea, 20 mg dithiothreitol, 20 μL Pharmalyte [3-10], 1.4 mg phenylmethylsulfonyl fluoride, 1 μg each aprotinin, leupeptin, pepstatin A, and antipain 50 μg TLCK, and 100 μg TPCK/1 mL). We applied 100 μL each MPRO cell lysate (2.5×10^6cells/100 μL) to immobilized pH gradient (IPG) strips (pH 3-10 L; Amersham Pharmacia Biotech), and IPG electrophoresis was conducted for 16 hours (20 100 Vh) using an Immobiline Drystrip Kit (Amersham Pharmacia Biotech). Electrophoresis in the second dimension was carried out in a 12% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) gel with the Laemmli-SDS continuous system in a Protean II xi 2-D cell (Bio-Rad) run at 40 mA constant current for 4.5 hours. Proteins were detected by Brilliant Blue G–colloidal staining.[24] Protein spots were excised from the gel and digested with trypsin. ACTH clip (average [M+H] 2466.70) and bradykinin (average [M+H] 1061.23) were used for calibration of peptide masses. One microliter sample digest was mixed with 1.0 μL α-cyano-4-hydroxy cinnamic acid (4.5 mg/mL in 50% CH_3CN, 0.05% TFA) matrix solution and 1 μL calibrants (100 fmol) each. The spectra of the peptides were acquired in reflector/delayed extraction mode on a Voyager-DE STR mass spectrometer (Perseptive Biosystems, Foster City, CA). Peptides were identified using the ProFound search engine.[39]

Results

Differentiation of MPRO cells

Figure 1 illustrates the morphologic changes in an MPRO cell population representative of those used for RNA expression analysis. Undifferentiated MPRO cells resembled promyelocytes under the light microscope (Figure 1A). After induction with ATRA for 24 hours, the cells morphologically differentiated into metamyelocytes (Figure 1B). At 48 hours, the cells further developed into metamyelocytes and band neutrophils (Figure 1C). At 72 hours, nearly 100% of MPRO cells became mature neutrophils (Figure 1D).

Identification of mRNAs by differential display assay

MPRO cellular mRNA was analyzed at 0, 24, 48, and 72 hours after ATRA treatment. Nine restriction enzymes were used in a 3′-end DD approach. During MPRO differentiation, 1109 fragments corresponding to 837 transcripts were found to change substantially in expression levels (Figure 2). These represented approximately 279 known genes, 112 ESTs, and 59 putative new genes, each with a perfect or fair polyadenylation signal at an appropriate distance from the oligo-dT priming site. The gene information detected by DD was collected in database dbMCd.

Identification of mRNAs by oligonucleotide chip assay

We used an oligonucleotide chip containing 13 179 probe sets corresponding to approximately 7000 murine genes to analyze patterns of mRNA expression in the same RNA samples used for DD. The information obtained by oligonucleotide arrays was collected in the database dbMCa.

We clustered the genes by their similarity to idealized expression patterns. For instance, the expression pattern of an ideal gene that is overexpressed (high) at time 0 and underexpressed (low) at 24, 48, and 72 hours, would be high-low-low-low (HLLL). Overall we have (2^4-2) idealized patterns excluding HHHH and LLLL. Pearson correlation was used as the measure of similarity of each gene expression pattern, $x = (x_1,x_2,x_3,x_4)$ to each of the 14 idealized patterns $y = (y_1,y_2,y_3,y_4)$. The 4 entries of x and y corresponded to the 4-dimensional gene expression levels at 0, 24, 48, and 72 hours, respectively. Each gene was assigned to a cluster labeled by the idealized pattern that had the maximal correlation with that gene. We selected only genes that hybridized well compared with the background (considered "present" by GeneChip software) and had maximal AD amplitude greater than 200 U in at least 1 of the 4 stages. We further tabulated the 14 patterns according to whether the gene expression changed at early (0-hour), intermediate (24- and 48-hour), and late (72-hour) time points and whether gene expression monotonically increased (up-regulated), monotonically decreased (down-regulated), or was not monotonic (transient). Table 1 shows 8 clusters of 104 genes that had significant changes of mRNA levels, arranged according to the temporal stage and the monotonic/transient changes of expression levels.

Principal component analysis determined whether we could comprehensively present multidimensional data (4-dimensional in our case) in a simple 2-dimensional graph. First, we found the 4 principal components, which were the axes of the most compact 4-dimensional ellipsoid that encompassed the 4-dimensional cloud of data. Each axis was a different linear combination of the original 4 variables. Then we verified that the first 2 principal components (the first 2 largest axes of the ellipsoid) captured most (95.2%) of the variation of the data. Therefore, the data could be faithfully projected (with a minor loss of information) into a 2-dimensional graph, with the 2 largest principal components as the x- and y-axes. As shown in Figure 3, genes tend to coalesce in clusters, according to their labels determined by their similarity to an ideal expression pattern. In summary, a genomic (global) picture of the distribution of genes according to their similarity to predetermined idealized multidimensional expression patterns is concisely displayed in a 2-dimensional graph.

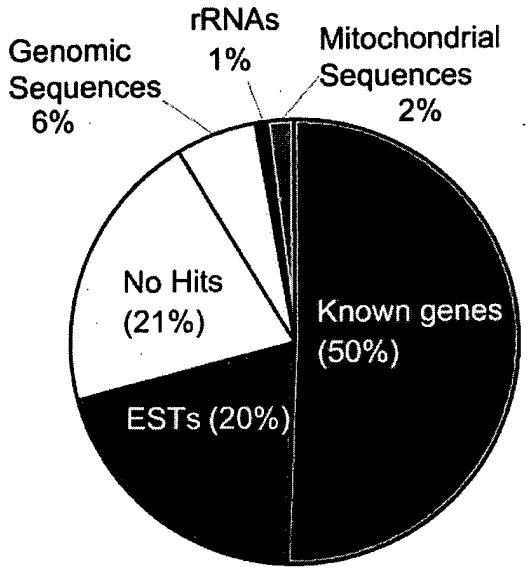


Figure 2. Distribution of genes obtained by DD assay. MPRO cell mRNA was analyzed at 0, 24, 48, and 72 hours after ATRA treatment; 1109 fragments corresponding to 837 transcripts were found to change substantially in expression levels. The total 837 transcripts were classified into 6 categories according to the bioinformatic analysis. Percentages show the gene distributions in these 6 categories. Information for each transcript was collected in database dbMCd.

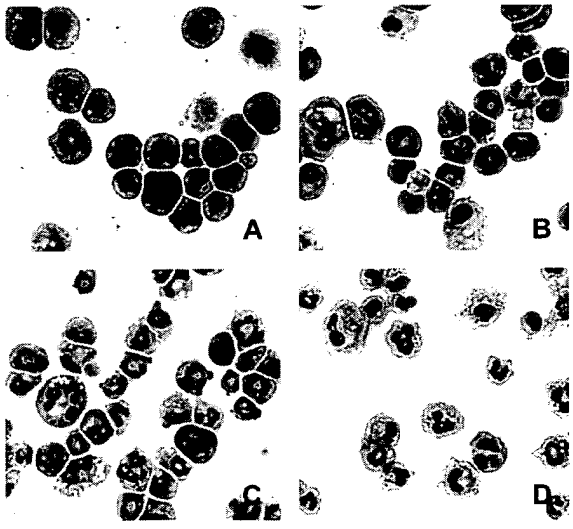


Figure 1. Morphology of MPRO cells during differentiation. MPRO cells were induced as described in "Materials and methods," concentrated by cytospin, and Wright-Giemsa stained. (A) Uninduced MPRO cells. (B) MPRO cells induced with ATRA for 24 hours. (C) MPRO cells induced with ATRA for 48 hours. (D) MPRO cells induced with ATRA for 72 hours.

Table 1. Genes differently regulated during the different stages of mouse promyelocytic cell line differentiation process

Category	Timing		
	Early	Middle	Late
Up-regulation	LHHH (n = 10) *Mad* P2rx1 **Itgb2 Il1r2 Lcn2** *Itpr5* Cebpb H2-D Etohi6 Zyx	LLHH (n = 6) *Piral Cybb* Pfc Pira5 Cd53 *Ifngr2*	LLLH (n = 13) *Il1a Csflr* Ii *Ctsl* **S100a8** L-CCR *Ctss* **Aldo1 Rac2** Fpr1 **Ctsd Ubb Ptmb4**
Down-regulation	**HLLL** (n = 11) Tcrg-V4 Ly64 Ctsg Spi2-1 Mcpt8 Myc Myb Tlr4 Npm1 Erh Hsp60	**HHLL** (n = 1) **Mpo**	**HHHL** (n = 37) Actx *Irf2* EL2 **Rpl19 Actb Ly6e** *Atf1* Hist2 Psma2 **Gnas** *Zfp36 Il4ra Ltbr* Shfdg1 Max **Rps8** Csf2rbl Slpi Tctex1 **Tpi** Btf3 *Cntf* Gys3 Slc10a1 Ctsb Sepp1 Rtn3 Ccnb2 S100a9 Cf11 Hist5-2ax Rela Copa Gstm1 **Gnb2-rs1** Grn RPL8
Transient		**LLHL** (n = 9) Sell *Klf2* Pira6 Pirb Lst1 Ltf Sema4d *Stat6* Mmp9 **LHHL** (n = 17) *Cebpa* **Lyzs Fcgr3** Arf5 Lamp1 Stat3 Csf2ra Osi **Actg** Sfpi1 Gpx3 Ptprc Prtn3 Irf1 Rps6ka1 Ltb4r Myln	

Arrays of Affymetrix Mu11k containing 13103 probe sets corresponding to 12002 GenBank accessions were used for hybridization. Arrays were hybridized with streptavidin-phycoerythrin (Molecular Probes) biotin-labeled RNA and scanned. Intensity for each feature of the array was captured using Genechip software (Affymetrix), and a single raw expression level for each gene was derived from the 20 probe pairs representing each gene using a trimmed mean algorithm. For each gene, an AD of 24-, 48-, and 72-hour samples was calibrated by dividing the slope of the linear regression line for a graph with the x-axis the AD of 0-hour probe sets and the y-axis the AD of the respective time point (24, 48, or 72 hours). A threshold of 20 U was assigned to any gene with a calculated expression level below 20 because discrimination of expression below this level could not be performed with confidence.[38] Each gene expression profile was categorized as described in Tables 3, 4, and 5. For the 4 time points, the minimum AD of the relatively higher group (MIN-H) was divided by the maximum AD of the relatively low group (MAX-L), and those genes whose MIN-H/MAX-L greater than 2 were selected as meaningfully regulated. Genes were sorted in descending order based on the MIN-H/MAX-L. Genes in boldface are those whose expression level was in the top 20% (ie, maximum AD of 4 time points greater than 3000), and genes in italics are those in the bottom 20% (ie, maximum AD of 4 time points less than 300). The differentiation period was grouped into 3 stages: early (0-hour), middle (24-hour and 48-hour), and late (72-hour) stages.

AD indicates average difference; gene symbols are expanded in an Appendix at the end of this article.

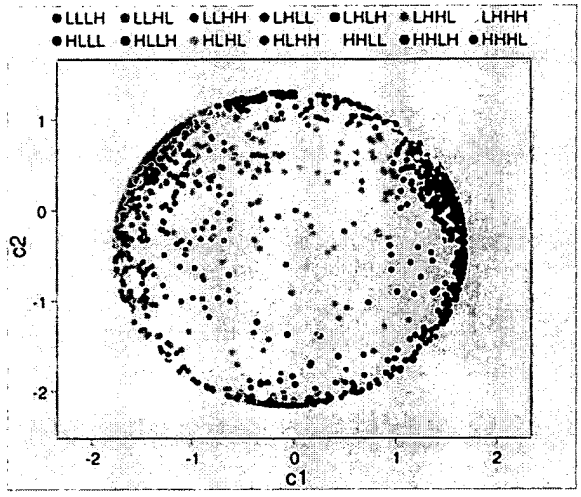


Figure 3. Gene clusters in the first 2 principal component spaces. Principal component analysis allowed us to present the multidimensional data (in this case, 4-dimensional data of each gene expression pattern) in a simple 2-dimensional graph. We derived the 4 principal components, which are a linear combination of the standardized expression intensities (zero mean and unit variance) at 0, 24, 48, and 72 hours. The first 2 principal components captured most of the variation of the data (approximately 85%). Therefore, the data can be displayed (with a minor loss of information) in a 2-dimensional graph. The first and second principal components, c1 and c2, are given by the linear combinations $c_1 = 0.747 \cdot n1 - 0.11 \cdot n2 - 0.656 \cdot n3 + 0 \cdot n4$ and $c_2 = 0.278 \cdot n1 + 0.353 \cdot n2 + 0.233 \cdot n3 - 0.863 \cdot n4$, where $n1$, $n2$, $n3$, and $n4$ are the rescaled and standardized expression levels at 0, 24, 48, and 72 hours, respectively. The axes legends c1 and c2 stand for the first 2 principal components. In this paper we used the Pearson correlation to measure the similarity of each gene with the idealized expression patterns, as opposed to the Euclidean distance we used in a previous work,[19] because clusters were better separated using this measure. In both cases, we presented the data in the 2-dimensional space of the lowest principal components. The data had a tendency to be circularly distributed when we used the Pearson correlation as a distance measure.

Correlation between array and DD analyses

We have previously demonstrated a correlation coefficient of 0.93 between visual estimates of changes in band intensity on DD and Phosphorimager System (Molecular Dynamics, Sunnyvale, CA) estimates of band intensity and a correlation coefficient of 0.88 between hybridization intensity changes of mRNA on Northern blot analyses and changes in band intensity on DD.[19] In a few cases there were clear discrepancies in the pattern of expression of a gene, as estimated by DD and by oligonucleotide chip analysis. We chose the 6 most extreme cases and examined the levels of mRNA change for these genes by Northern blot analysis (Figure 4). In 5 cases, the Northern blot results agreed with the results of the DD analysis, whereas the results of Gnb2-rs1 disagreed with the oligonucleotide array but duplicate bands from DD showed a relatively high level of expression in the 0 time sample that did not correlate with the Northern blot (Table 2). One possible explanation for these findings was the change in the relative use of different polyadenylation sites after the addition of ATRA to the MPRO cells.

Constructing a database for mRNA level changes during myeloid differentiation

Based on the data obtained above, an in-house database (dbMC) was constructed that included 2 subdatabases, dbMCd and dbMCa, for collecting gene information from DD or oligonucleotide arrays, respectively. Each entry in dbMC is accompanied by a so-called executive summary. The linkage between dbMCd and dbMCa was established by UniGene ID and cluster ID. dbMC contains the temporal expression patterns of genes during the MPRO cell differentiation process, including not only products represented in public databases but also novel transcripts.

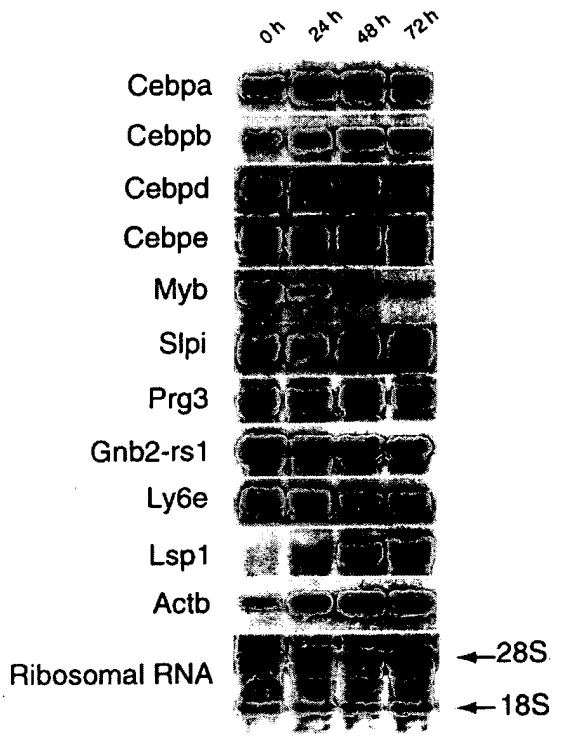


Figure 4. Northern blot analysis of selected mRNAs. Equivalent amounts of RNA from MPRO cells induced by ATRA at different time points (0 hour, 24 hours, 48 hours, and 72 hours) were resolved by formaldehyde-agarose gel electrophoresis, stained to verify the amount of loading. Eleven genes were separately probed on the RNA filters. The gene symbol of each probe was listed at the left of a related Northern blot result. Detailed information on these 11 probes was listed in Table 5. One of the RNA-blotted membrane photographs is shown with methylene blue–stained 28S and 18S RNA subunits demonstrating the quality and quantity of RNA loaded in individual lanes.

Analysis of gene expression patterns during MPRO differentiation

Many of the genes identified in this study were found in myeloid cells or were implicated in myeloid development for the first time. We detected 8 cytokines[25] and chemokines whose mRNA levels changed more than 5-fold by arrays and 2-fold by DD during the maturation of MPRO cells (see our Web site, http://bioinfo.mbb.yale.edu/expression/neutrophil). Among these were 2 members of the CC chemokine family. Interleukin-1α (IL-1α) was up-regulated at the late stage of differentiation (LLLH pattern, Table 1).

mRNA for approximately 52 receptors was detected by one or the other method. A number of the receptors known to be present on mature neutrophils showed late induction of mRNA, and their levels of induction were high, indicating that the expression of these products is a prominent event late in neutrophil maturation (Table 3). Rarely was mRNA for receptors down-regulated, consistent with myeloid maturation being accompanied by increasing responsiveness of the cell to a variety of external stimuli.

Expression of mRNA for granule proteins

Neutrophils contain several types of granules that develop at different stages of myeloid maturation.[3,17,26] Levels of mRNAs encoding secondary granule proteins, such as lactoferrin, increased as the cells matured (Table 4). The level of mRNA for Mmp9, reported as a tertiary granule protein, increased markedly between 24 and 48 hours after the induction of differentiation, whereas mRNAs for secondary granule proteins either increased less markedly or showed a maximum increase by 24 hours. mRNAs for several primary granule constituents, such as myeloperoxidase and cathepsin G, were present in unstimulated cells and decreased as the cells matured. There was a discrepancy in the measurements of proteoglycan mRNA by DD and oligonucleotide chips, but Northern blots showed that it reached a peak at 48 hours and then declined (Figure 4). Cathepsin D is reported as a primary granule protein, but its pattern of mRNA expression more closely resembled that of secondary granule constituents. In addition to known granule components, mRNAs for several other cathepsins were up-regulated during myeloid differentiation, in parallel with or later than the tertiary granule protein mRNAs.

mRNAs for transcription factors

Transcription factor genes, including several identified at the sites of consistent chromosome rearrangements in acute myeloid leukemia, have been implicated in normal myeloid differentiation and in the expression of neutrophil proteins.[2,5,27] However comprehensive information concerning the expression of these transcription factors during myeloid development is not readily available. Therefore, we compared gene names and identifiers in our databases to those of the transcription factor database Transfac (http://

Table 2. Expression patterns of genes detected by Northern blot analysis

Gene symbol	Gene accession	AD value by array				Intensity by DD			
		0 h	24 h	48 h	72 h	0 h	24 h	48 h	72 h
Cebpa	M62362	33	212	182	44	—	—	—	—
Cebpb	X62600	390	1248	1380	1903	—	—	—	—
Cebpd	X61800	157	262	168	430	—	—	—	—
Cebpe	—	—	—	—	—	—	—	—	—
Myb	M12848	892	356	230	435	—	—	—	—
Slpi	U73004	617	501	783	402	**1**	**2**	**3**	**3**
Prg3	W45834	153	259	339	345	**5**	**1**	**1**	**2**
Gnb2-rs1	X75313	4231	3623	3215	3403	**4**	**4**	**1**	**1**
Ly6e	U04268	3061	5391	2844	1282	**3**	**2**	**1**	**1**
Lsp1	M90316	65	376	840	28	**2**	**3**	**5**	**6**
Actb	X03765	3095	3588	3976	2434	**1**	**2**	**3**	**2**

Gene symbol and gene accession refer to National Center for Biotechnology Information databases and, in particular, to Locus Link. AD value is the average difference in the value of hybridization intensity between the set of perfectly matched oligonucleotides and the set of mismatched oligonucleotide in the oligonucleotide array. Band intensities from DD were semiquantified on a scale from 1 (+) to 8 (++++++++). These estimates are shown as boldface numbers in this table.[19] Both AD value and intensity of genes were studied at 4 time points corresponding to MPRO cells induced for the indicated times.

DD indicates differential display; MPRO, mouse promyelocytic cell line; for gene symbols, see the Appendix at the end of this article.

Table 3. Receptors expressed during myeloid differentiation process

Maximal fold change	Gene symbol	Gene accession	AD value by array			
			0 h	24 h	48 h	72 h
Less than 2						
	Bzrp	D21207	641	658	881	887
	Cmkar4	X99581	508	447	378	684
	Crry	M34173	433	384	506	506
	Csf2rb1	M34397	318	345	410	241
	Htr5a	Z18278	188	272	273	339
	M6pr	X64068	536	409	408	649
	MPPIR	AA116789	232	84	63	381
	TCRGB	M26053	165	212	244	299
	Tnfrsfla	M59377	**0**	**1**	**1**	**1**
2 or more, less than 3						
	Cmkbr1	U28404	221	244	504	638
	Crhr	X72305	121	200	250	355
	Csf2ra	M85078	171	372	402	254
	Ebi3	AF013114	187	270	428	148
	Grid1	D10171	128	164	150	257
	Ifngr	J05265	141	263	327	251
	Il2rg	U21795	205	184	231	477
	Ldlr	X64414	1399	1653	1665	3968
	P40-8	J02870	849	677	381	640
	Plaur	X62701	312	443	476	734
	Rarg	M34476	102	113	114	218
	Srb1	U37799	126	232	132	258
3 or more, less than 4						
	Cr2	M29281	83	138	243	77
	Csf2rb2	M29855	209	249	437	111
	Fcer1g	J05020	2398	2766	3365	8751
	Fcgr2b	X04648	1703	1652	1431	4605
	Ifngr2	U69599	**1**	**2**	**2**	**3**
4 or more, less than 5						
	Nr4a1	X16995	96	188	202	401
5 or more						
	Il1r2	X59769	482	1796	2872	3818
	C5r1	L05630	185	434	808	1078
	Drd2	X55674	0	0	0	219
	Fcgr3	M14215	**1**	**1**	**1**	**2**
	Fpr1	L22181	0	89	141	671
	GCR	AA240711	**2**	**0**	**0**	**0**
	L-CCR	AA034646	48	175	314	2056
	NMDARGB	AAB20211	**2**	**2**	**0**	**0**
	P2rx1	X84896	79	346	530	744
	Pira1	U96682	0	43	172	378
	Pira5	U96686	274	391	954	1874
	Pira6	U96687	122	635	2014	1716
	Pirb	U96689	191	445	966	747
	Sell	M25324	46	104	570	20
	Tcrg-V4	M54996	1650	78	65	315

Receptors are identified as present whose maximal AD values were more than or equal to 200 U in this study. Genes were sorted by their expression patterns as follows: first by the average difference value, then by the difference between minimum and maximum AD for the 4 time points, and last by the alphabetical order of gene symbols. Genes were ordered according to the maximal fold change of AD values. Abbreviations of gene names are taken from gene symbols listed in the Locus Link portion of the National Center for Biotechnology Information database where available. Numbers in bold denote those gene expression patterns obtained by differential display rather than by oligonucleotide array assays. The other information is presented as in the legend to Table 2.

AD indicates average difference; gene symbols are expanded in an Appendix at the end of this article.

www.transfac.gbf-braunschweig.de/TRANSFAC) and determined which factors contained in this database were present at detectable levels in MPRO cell mRNA, using Affymetrix software for the criteria for inclusion of mRNAs from approximately 200 murine transcription factors probe sets on the oligonucleotide chip. Of these, 54 were expressed and 13 showed changes of 3-fold or more in chip signal (Table 5).

The changes in certain transcription factors, such as the moderate down-regulation of *myb* and *myc* and the up-regulation of the Max dimerization protein MAD, were consistent with the shift of the cells from a proliferative to a differentiated state.[28] Some changes are more difficult to explain, such as the up-regulation of DP1, a partner for E2f factors in the regulation of S-phase genes, and the mild up-regulation of the *Id* genes, commonly associated with an inhibition of differentiation by competition with bHLH transcriptional activators.[29]

The C/EBP family has been extensively studied with respect to myeloid differentiation.[2,30] Absolute levels of the C/EBP α and δ mRNAs were low, probably at the borderline of significance for the oligonucleotide chip assay, whereas the level of C/EBP β appeared higher. In addition, there were discrepancies between the chip

Table 4. Granule constituents expressed during mouse promyelocytic cell line cell differentiation

Granule constituent	Gene symbol	Gene accession	AD value by array			
			0 h	24 h	48 h	72 h
Azurophil (primary) granules						
	Man2c1	AA161860	178	134	99	164
	Ctsb	M65270	442	480	595	389
	Ctsd	X52886	214	1087	1828	2784
	Ctsg	M96801	1509	405	46	286
	E12	U04962	658	1273	843	157
	E1a2	AA689016	47	159	134	163
	Gus-s	M63836	544	226	266	254
	Lyzs	M21050	0	1	1	3
	Mcpt8	X78545	831	268	66	491
	Mpo	X15378	3788	3009	776	692
	Prg	X16133	2621	2653	2920	9859
Possible granule proteins						
	Ctsc	AA144887	252	194	342	576
	Ctse	X97399	1	3	4	5
	Ctsh	U06119	45	124	195	156
	Ctsl	X06086	16	11	31	237
	Ctss	AA089333	12	9	88	463
Specific secondary granules						
	Cpa3	J05118	621	270	90	801
	Cd36l2	AB008553	113	93	157	187
	Cnlp	X94353	80	479	704	626
	Cybb	U43384	8	24	91	128
	Ear2	—	0	1	1	2
	Fpr1	L22181	178	220	235	846
	Itgb2	X14951	0	2	4	2
	Lcn2	W13166	916	3513	3931	6036
	Ltf	J03298	19	162	333	138
	MBP	W45834	5	1	1	2
	Mmp13	X66473	44	43	72	178
	Ngp	L37297	2661	4782	2311	6912
Tertiary granules						
	Mmp9	Z27231	0	1	2	2

Shown are the possible granule protein cDNAs represented on the oligionucleotide arrays, sorted by their expression patterns as follows: first by the average difference AD value, then by the granule types, and last by the alphabetical order of gene symbols. Data are presented as described in the legend to Table 3.

AD indicates average difference; gene symbols are expanded in an Appendix at the end of this article.

estimates and the mRNA levels observed by Northern blotting with specific probes for these genes. In particular, the latter method, more sensitive and specific, showed that C/EBP α began to decline in the most mature cells, whereas C/EBP δ mRNA declined progressively beginning at 24 hours after the onset of differentiation.

C/EBP ε is a more recently cloned C/EBP family member. Previous studies indicated it is expressed in a large array of human leukemia cell lines blocked at various stages of differentiation and that it is up-regulated during granulocytic differentiation.[31] A C/EBP ε probe was not included in the oligonucleotide chips, and this mRNA was not detected by DD. Therefore, we examined the C/EBP ε expression patterns by quantitative PCR and Northern blot analysis (Figure 4). C/EBP ε exon 1 was PCR amplified from MPRO RNAs using primers RY48 (AGCCCCCGACACCCTTGATGA) and RY49 (TGGCACACT-GCGGGCAGACAG).[32] The results showed that C/EBP ε is expressed throughout myeloid differentiation, with expression levels increased moderately in the later stages.

We detected a number of other transcription factors that are broadly expressed or that have been reported in other studies of hematopoiesis (Table 5). Some of the factors that were most strongly induced during differentiation have been studied in other contexts but not previously implicated in hematopoiesis, such as a mammalian homologue to the *Drosophila* enhancer of split gene, a transcriptional silencer. The mammalian gene is expressed at relatively high levels as measured by the oligonucleotide chip and is a candidate for mediation of the silencing of growth-related genes in the maturing neutrophil. Another candidate transcriptional silencer, Tif1b, may serve as a corepressor for the KRAB domain family of zinc finger transcription factors and also may mediate binding of the heterochromatin protein HP1 to DNA.[33]

There were 26 transcription factors whose mRNAs showed no significant changes by oligonucleotide chip analysis and were not identified as differentially regulated genes by differential display assays. PU.1, a factor necessary for the production of neutrophils and the expression of several neutrophil genes,[34] showed less than a 3-fold increase in mRNA, below the threshold for a significant change. Other candidate hematopoietic transcription factors, such as PEBP1aB2 (AML1), GATA-1, and SP-2, were represented on the oligonucleotide chips, but their mRNA levels were so low that they were reported as absent in this study. The possibility that small changes in the levels or ratios of some transcription factors could produce marked changes in transcription potentially limits the ability of data generated by present methods to explain transcriptional changes during differentiation.

Protein expression patterns of MPRO cells during ATRA induction

We visually compared the 2DE patterns from MPRO cells at the same time points used for mRNA analysis. In most cases the

Table 5. Transcription modulators presented during myeloid differentiation

Maximal fold change	Gene symbol	Gene accession	AD value by array			
			0 h	24 h	48 h	72 h
Less than 2-fold						
	Zfp11-6	AB020542	2630	2989	2795	2515
	Btf3	W13502	3	3	2	1
	Gata2	AB000096	562	770	472	730
	Hmgi	J04179	337	348	177	232
	Idb1	M31885	455	787	721	637
	Max	M63903	256	224	312	172
	Nfatc2	AA560093	2313	3218	2396	2542
	Pm1	U33626	173	281	329	306
	Rarg	M34476	102	113	114	218
	Rela	M61909	297	260	304	244
	Sox15	W53527	419	461	484	837
	Ybx1	M62867	643	489	472	496
	Zfp162	Y12838	671	734	720	992
2 or more, less than 3						
	Cebpd	X61800	157	262	168	430
	Idb2	M69293	244	210	310	604
	Jund1	W29356	1274	2002	1434	3085
	Lyl1	X57687	399	342	347	891
	Nfe2	L09600	458	743	1042	505
	Nfkb1	L28117	953	2044	1876	2034
	Pbx1	AF020196	611	303	345	212
	sfpi1	A34693	375	784	991	529
	Tif1b	U67303	673	659	420	863
	Trp53	P10361	259	149	125	361
	Usf2	U12283	129	185	285	192
	Ybx3	L35549	96	169	210	119
	Zfp216	AA510137	82	151	204	106
3 or more, less than 4						
	Irf1	M21065	85	207	278	198
	Klf2	U25096	62	86	246	77
	Myb	M12848	892	356	230	435
	Stat3	AA396029	484	1057	1012	290
	Tfdp1	Q08639	307	560	505	1093
4 or more, less than 5						
	Cebpb	X62600	390	1248	1380	1903
	Stra14	Y07836	223	383	510	936
5 or more						
	Cebpa	M62362	33	212	182	44
	Grg	X73359	99	565	916	1005
	Mad	X83106	0	111	167	327
	Myc	L00039	314	112	62	173
	Etohi6	W89667	169	386	313	1003
	TBX1	AA542220	0	0	1	2

Shown are the transcription factors identified as present by the oligonucleotide array analysis whose maximal AD between perfect match and mismatch oligonucleotide sets was greater than or equal to 200 U in this study. Data are presented as described in the legend to Table 3.

AD indicates average difference; gene symbols are expanded in an Appendix at the end of this article.

peptides identified for a given protein were derived from regions along the entire length of the protein, indicating the observed products were not the result of proteolytic degradation. These data must be considered with several caveats: membrane and other hydrophobic proteins and very basic proteins are not well displayed by the standard 2DE approach, and proteins present at low levels will be missed.[35] In addition, to simplify MS analysis, we used a Coomassie dye stain rather than silver to visualize proteins, and this decreased the sensitivity of detection of minor proteins. The MS method we used was sufficiently sensitive to identify proteins that could barely be visualized by colloidal blue staining. However, a limitation of the method for the mouse is that the current database lacks predicted amino acid sequences for a substantial fraction of murine genes. In addition, very small proteins give only a few peptides, making statistically confident identification difficult.

Figure 5 shows the analytical colloidal blue–stained 2DE IPG reference maps of differentiated MPRO cells. Expression patterns of more than 500 protein spots were detected and observed through the entire series of gels. Protein spots could easily be cross-matched to each other, indicating the reproducibility of the method. As marked on the gel pictures (Figure 5), 50 proteins with a wide range of molecular weights (1 to 200 kd), isoelectric points (4 to 9), and abundances were subjected to MS protein identification. The results are presented in Table 6.

Comparing the theoretical value of the molecular weight and *pI* of each protein to that of the observed value, we confidently identified 28 proteins in the expected position on the gels (spots 1 to 28). Some of the other proteins with strong matches to the murine databases migrated to a somewhat unexpected *pI* position. Nine spots gave clear peptide peaks on mass spectroscopy but did not match any known gene. Their identification will require amino acid

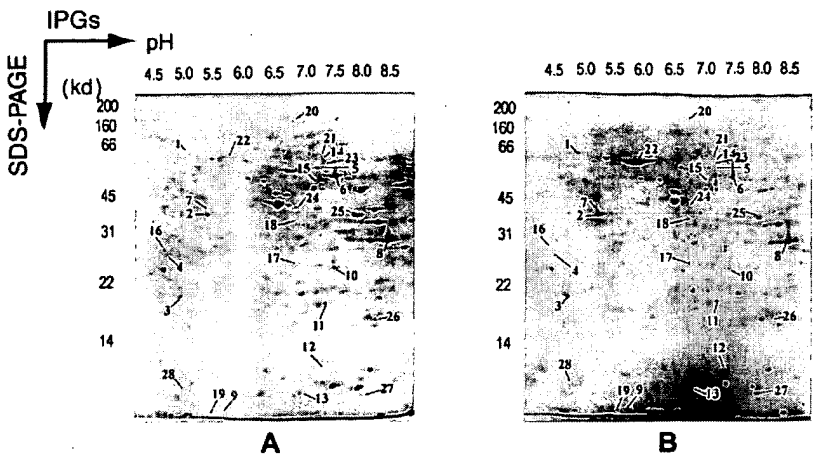


Figure 5. 2DE electrophoretograms of MPRO cells. MPRO cell lysate (2.5×10^6 cell/sample) was loaded for 2DE analysis. Gels were stained with brilliant blue G–colloidal dye. (A) 2DE map of uninduced MPRO cell (0 hour). (B) 2DE map of matured MPRO cells (72 hours). Protein spots marked in the maps were considered differentially expressed and were subjected to MS analysis. The resultant protein information is listed in Table 6.

sequence analysis or availability of more extensive murine databases. We searched for the expression patterns of the genes cognate to the expressed proteins in dbMC (Table 6). Nineteen genes were found in dbMC, the mRNA for 5 genes was reported as absent, and 13 genes were present during MPRO differentiation. Comparison of the expression patterns showed only 4 genes of 18 present on the oligonucleotide chips whose expression was consistent at the RNA level and protein level. None of these was on the list of the genes that were differentially expressed significantly (5-fold or greater change by array or 2-fold or greater change by DD).

Discussion

We explored the temporal patterns of gene expression during myeloid development. A database has been developed to provide a

Table 6. Correlation of expression patterns between mRNA level and protein level

Spot	Protein definition	Gi number	Predicted value kd	Predicted value *pI*	Percentage (%)	2DE pattern 0 h	2DE pattern 72 h	cDNA expression pattern 0 h	cDNA expression pattern 72 h	Ag
1	GRP 78	2506545	72.4	5.1		1	3	1321	1043.3	N
2	Actin, gamma, cytoplasmic	6752954	41.77	5.3	40	3	6	**0**	**2**	Y
3	RHO GDI 2	2494703	22.83	4.9	33	3	3	341	441.6	Y
4	Proliferating cell nuclear antigen	7242171	28.77	4.7	42	1	0	544	430.9	Y
5	APS kinase	4038346	69.8	7.1	24	2	1	43	50.7	N
6	Pyruvate kinase 3	6755074	57.9	7.2	48	6	4	3047	5880.3	N
7	Melanoma X-actin	6671509	41.72	5.3	39	1	3	2539	341.3	N
8	Glyceraldehyde-3-phosphate dehydrogenase	6679937	35.79	8.7	39	8	7	3073	5742.3	N
9	Stefin 3	461911	10.99	5.9	48	0	4	N/A	N/A	—
10	Guanine nucleotide binding protein, beta-2, related sequence1	6680047	35.06	7.9	21	4	2	139	303.1	N
11	Triosephosphate isomerase	6678413	26.69	6.9	26	3	3	3312	2660.1	Y
12	Testis-derived c-abl protein	1196524	17.19	7	51	2	3	152	126.9	N
13	RNA binding motif protein 3	7949121	16.59	6.8	25	1	0	628	812.4	N
14	Collapsin response mediator	6681019	62.16	6.4	36	2	0	Absent	Absent	N
15	Lamin A	220474	47.52	6.6	35	2	0	Absent	Absent	N
16	47-kd keratin	52783	35.82	4.8	29	3	0	Absent	Absent	N
17	sid478p	5931565	31.3	6.7	30	1	2	Absent	Absent	N
18	MHC class II H2-IA-beta-5	3169662	28.6	7.1	39	1	2	N/A	N/A	—
19	Androgen-binding protein: subunit alpha	739346	8.04	6.4	68	0	2	Absent	Absent	N
20	Neuronal apoptosis inhibitory protein	5932010	158.7	6	17	1	0	N/A	N/A	—
21	PAD type IV	6755018	74.46	7.2	21	1	3	N/A	N/A	—
22	Human serum albumin homologoue	3212625	66.45	5.7	24	0	6	N/A	N/A	—
23	syncrip	6576815	62.53	7.2	33	2	1	N/A	N/A	—
24	Transamidinase	1730203	48.22	7.2	31	3	1	N/A	N/A	—
25	PGK crigr phosphoglycerate	1730519	44.54	8.3	47	5	4	1088	1402.3	N
26	Proliferation-associated gene A	6754976	22.16	8.6	53	3	1	N/A	N/A	—
27	Putatuve peroxisomal antioxidant enzyme	3913065	17	7.8	55	0	3	N/A	N/A	—
28	IgE chain C2 region	2137430	12.1	5.2	38	0	1	N/A	N/A	—

The proteins listed here are represented by the spots marked in the electrophoretograms shown in Figure 5.

Protein definition, Gi number, and predicted value refer to the protein name, accession number, and properties derived from the National Center for Biotechnology Information protein database. The column labeled % shows the percentage of peptides predicted from the protein sequence that were detected by mass spectroscopy. The expression level of protein spots expressed in mouse promyelocytic cell line cell induced by all-*trans* retinoic acid for 0 hours and 72 hours (Figure 5) were scored on a scale of 1 (+) to 8 (++++++++) in the 2DE pattern column. The cDNA expression patterns of the cognate mRNAs are listed in the cDNA expression pattern column abstracted from the dbMC database. The genes not represented on the oligonucleotide arrays were marked as N/A. Ag showed the correlation of gene patterns at mRNA level or protein level.

Y indicates agreement and N discrepancy between changes in cDNA and protein spot intensity. The numbers in bold were obtained with DD. 2DE indicates 2-dimensional gel electrophoresis; IgE, immunoglobulin E; DD, differential display.

reference for later research on the molecular mechanisms underlying normal myeloid development.

The MPRO cell system morphologically mimics normal myeloid differentiation and biochemically proceeds further toward mature neutrophils than most other in vitro systems. Because the arrest in differentiation of MPRO cells growing in the absence of ATRA is not physiologic, there is a theoretical risk that gene expression in these cells is not coordinated in the way that it is in normal differentiation. It is encouraging that, for the most part, the timing of expression of genes for proteins of the various neutrophil granules is consistent with the timing of the morphologic and biochemical appearance of these granule components during normal myeloid differentiation.

The DD technique provides certain advantages for detecting and comparing mRNA levels in different samples. First, the method is, in principle, similar to competitive RT-PCR, and, with the use of stringent PCR conditions, is expected to be about as reliable. Second, display patterns are reproducible. Third, the method detects the levels not only of RNAs already represented in the database but also of unknown RNA species that may represent "new" genes. Fourth, closely related genes can be distinguished regardless of cross-hybridization, provided there are some single nucleotide differences in the 3′ end sequence. Limitations associated with this technique are that numerous gels are necessary to get complete information and that comparison of the levels of different mRNAs is only approximate because of the differential amplification of bands of different size or sequence.

Oligonucleotide chip analysis is a fast and effective means of accessing mRNA expression patterns.[20] Cluster analysis of groups of samples by this approach is effective. However, the present results indicate that alternative methods of verification are desirable before the data on an unexpected change in a particular gene are definitively accepted.

To obtain the broadest range of information from the myeloid differentiation process, both differential display and oligonucleotide chip techniques were applied in the current study. As a result, 65.3% of the observed changes in mRNA levels came from the differential display method and 41.5% came from oligonucleotide chip assays.

Our data showed in general that changes in expression pattern by the 2 methods agreed qualitatively but that there was some quantitative variation. Our results indicate that DD may be a more accurate way to detect changes in levels of gene expression than the oligonucleotide chip assay. However, improvements in the types of oligonucleotides used in arrays may close this gap in the future.

The mRNAs for a limited number of transcription factors vary in a pattern correlating with that of the mRNAs for primary or secondary granule proteins. However, more detailed information is needed, and the underlying mechanisms of granule gene regulation remain unclear. The number of potential positive and negative regulatory factors found here is sufficiently small as to make it feasible to perform in vivo studies, such as chromatin immunoprecipitation.

The oligonucleotide chip used in this study focused on known genes, whereas the DD method samples all polyadenylated transcripts. The latter method generated a large number of products not associated with known genes, in part because the mouse genome is not as well represented in the database as the human genome. However, our experience with DD and human mRNAs indicates that substantial fractions of the products represented as ESTs or not represented at all in the public databases are cDNA copies from introns, hnRNA, or other RNA with internal A runs.

Approximately 59 sequences obtained from gel-display bands had significant changes in the level of expression and a sequence that did not match that for any named gene in the public databases. Of these, 38 had plausible or excellent polyA signals. This is only an approximate estimate of the number of new genes found[36] because a fraction of the mRNAs for known genes still had poor polyA signals. In addition, the full 3′ untranslated region is often not known for characterized genes, and in some cases these new genes may prove to be identical to products identified by the oligonucleotide chips when more complete sequences are obtained. At the least, their presence indicates that a substantial fraction of the regulatory or functional circuitry of maturing myeloid cells remains unexplored and that valuable tools for their investigation will emerge from a combination of RNA expression studies and analysis of emerging genomic sequences.

The desired end point for the description of gene expression in a biologic system is not only the analysis of mRNA transcript levels but also the accurate measurement of protein abundance. The developments in 2DE and new MS instrumentation make it possible to accomplish this work rapidly and efficiently. In this study, we attempted to identify a number of the proteins differentially expressed between uninduced and ATRA-differentiated MPRO cells and to examine the relation between mRNA and protein expression levels for these genes representing the same state.

For protein levels based on estimated intensity of Coomassie dye staining in 2DE, there was poor correlation between changes in mRNA levels and estimated protein levels. Other groups have studied the correlation between mRNA and protein levels in yeast and liver cells.[11,12,14] In the liver cell experiments,[11,12] correlation coefficients of 0.4 to less than 0.5 were observed. In an extensive study in yeast,[11,12] the correlation coefficient was high if the most abundant mRNAs and proteins were considered. If a handful of these products was omitted, the remaining correlation coefficient was 0.4 or less. However, one could restore some of the correlation by averaging individual data points into broad proteomic categories.[37]

The discrepancies between mRNA and protein levels in MPRO cells appear to be substantially larger than those observed for yeast. Possible causes for the discrepancies include translational regulation, differential expression of certain mRNAs at various stages of cell growth in vitro, post-translational protein modification that varies with the stage of maturation of the cells, and selective degradation or excretion of proteins in vivo. Furthermore, here we are focusing on a developmental time-course, whereas the yeast study concentrated on the organism in vegetative growth. New techniques, equipment, and bioinformatic analysis tools must be developed to make such systematic, global, and quantitative analyses feasible.

The initial studies of protein expression presented here provide a cautionary note for efforts to interpret cell composition and function in relation to mRNA levels. Discrepancies we observed between gene expression and protein abundance suggest that selective post-transcriptional controls may be at least as important as changes in mRNA levels in determining the protein composition of neutrophils and that they are phenomena less well explored than transcriptional control. Analysis of mRNA expression patterns is itself only a small beginning toward a genome-wide description of cellular components.

Acknowledgments

We thank Dr S. Tsai (Fred Hutchinson Cancer Research Center) for his kind gift of the MPRO cell line, Dr Fuki M. Hisama (Yale University School of Medicine) for helpful advice, and the staff at Gene Logic Inc for data and support.

References

1. Lawson ND, Berliner N. Neutrophil maturation and the role of retinoic acid. Exp Hematol. 1999; 27:1355-1367.
2. Tenen DG, Hromas R, Licht JD, Zhang DE. Transcription factors, normal myeloid development, and leukemia. Blood. 1997;90:489-519.
3. Sigurdsson F, Khanna-Gupta A, Lawson N, Berliner N. Control of late neutrophil-specific gene expression: insights into regulation of myeloid differentiation. Semin Hematol. 1997;34:303-310.
4. Lenny N, Westendorf JJ, Hiebert SW. Transcriptional regulation during myelopoiesis. Mol Biol Rep. 1997;24:157-168.
5. Yunis JJ, Tanzer J. Molecular mechanisms of hematologic malignancies. Crit Rev Oncog. 1993;4: 161-190.
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270:484-487.
7. Stollberg J, Urschitz J, Urban Z, Boyd CD. A quantitative evaluation of SAGE. Genome Res. 2000;10:1241-1248.
8. Subrahmanyam YV, Baskaran N, Newburger PE, Weissman SM. A modified method for the display of 3′-end restriction fragments of cDNAs: molecular profiling of gene expression in neutrophils. Methods Enzymol. 1999;303:272-297.
9. Subrahmanyam YVBK, Yamaga S, Newburger PE, Weissman SM. A modified approach for the efficient display of 3′ end restriction fragments of cDNAs. In: Leslie RA, Robertson HA, eds. Differential Display: A Practical Approach. Practical Approach Series. Oxford, UK: Oxford University Press. 2000;101-129.
10. Appella E, Arnott D, Sakaguchi K, Wirth PJ. Proteome mapping by two-dimensional polyacrylamide gel electrophoresis in combination with mass spectrometric protein sequence analysis. EXS. 2000;88:1-27.
11. Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis. 1998;19:1853-1861.
12. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999;19:1720-1730.
13. The Whitehead Institute for Biomedical Research/MIT Center for Genome Research. Molecular Pattern Recognition Web site. Available at: www.genome.wi.mit.edu/MPR/analysis.html#RS. Accessed May 4, 2001.
14. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. Mol Cell Biol. 1999;19:7357-7368.
15. Tsai S, Collins SJ. A dominant negative retinoic acid receptor blocks neutrophil differentiation at the promyelocyte stage. Proc Natl Acad Sci U S A. 1993;90:7153-7157.
16. Johnson M, Caiazzo T, Molina JM, Donahue R, Groopman J. Inhibition of bone marrow myelopoiesis and erythropoiesis in vitro by anti-retroviral nucleoside derivatives. Br J Haematol. 1988;70:137-141.
17. Lawson ND, Krause DS, Berliner N. Normal neutrophil differentiation and secondary granule gene expression in the EML and MPRO cell lines. Exp Hematol. 1998;26:1178-1185.
18. Prashar Y, Weissman SM. Analysis of differential gene expression by display of 3′ end restriction fragments of cDNAs. Proc Natl Acad Sci U S A. 1996;93:659-663.
19. Subrahmanyam YVBK, Yamaga S, Prashar Y, et al. RNA expression patterns change dramatically in human neutrophils exposed to bacteria. Blood. 2001;97:2456-2468.
20. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science. 1996;274:610-614.
21. Lipshutz RJ, Chee M, Hubbell E, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques. 1995;19:442-447.
22. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High-density synthetic oligonucleotide arrays. Nat Genet. 1999;21:20-24.
23. Mao M, Fu G, Wu JS, et al. Identification of genes expressed in human CD34(+) hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. Proc Natl Acad Sci U S A. 1998;95:8175-8180.
24. Neuhoff V, Arold N, Taube D, Ehrhardt W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. Electrophoresis. 1988;9:255-262.
25. Wang Q, Miyakawa Y, Fox N, Kaushansky K. Interferon-alpha directly represses megakaryopoiesis by inhibiting thrombopoietin-induced signaling through induction of SOCS-1. Blood. 2000;96:2093-2099.
26. Gullberg U, Bengtsson N, Bulow E, et al. Processing and targeting of granule proteins in human neutrophils. J Immunol Methods. 1999;232:201-210.
27. Nichols J, Nimer SD. Transcription factors, translocations, and leukemia. Blood. 1992;80:2953-2963.
28. Amati B, Land H. Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. Curr Opin Genet Dev. 1994;4:102-108.
29. Pagliuca A, Gallo P, De Luca P, Lania L. Class A helix-loop-helix proteins are positive regulators of several cyclin-dependent kinase inhibitors' promoter activity and negatively affect cell growth. Cancer Res. 2000;60:1376-1382.
30. Yamanaka R, Lekstrom-Himes J, Barlow C, Wynshaw-Boris A, Xanthopoulos KG. CCAAT/enhancer binding proteins are critical components of the transcriptional regulation of hematopoiesis (review). Int J Mol Med. 1998;1:213-221.
31. Morosetti R, Park DJ, Chumakov AM, et al. A novel, myeloid transcription actor, C/EBP epsilon, is up-regulated during granulocytic, but not monocytic, differentiation. Blood. 1997;90:2591-2600.
32. Yamanaka R, Barlow C, Lekstrom-Himes J, et al. Impaired granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-deficient mice. Proc Natl Acad Sci U S A. 1997;94:13187-13192.
33. Nielsen AL, Ortiz JA, You J, et al. Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. EMBO J. 1999;18:6385-6395.
34. Anderson KL, Smith KA, Perkin H, et al. PU.1 and the granulocyte- and macrophage colony-stimulating factor receptors play distinct roles in late-stage myeloid cell differentiation. Blood. 1999;94:2310-2318.
35. Gorg A, Obermaier C, Boguth G, Weiss W. Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures. Electrophoresis. 1999;20:712-717.
36. Wilusz J, Pettine SM, Shenk T. Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. Nucleic Acids Res. 1989;17:3899-3908.
37. Jansen R, Gerstein M. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res. 2000;28:1481-1488.
38. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 1999; 96:2907-2912.
39. ProteoMetrics. ProFound search engine Web site. Available at: http://www.proteometrics.com/profound_bin/WebProFound.exe. Accessed May 4, 2001.

Appendix

Gene symbols used in tables: Actb: actin, beta, cytoplasmic; Actg: actin, gamma, cytoplasmic; Actx: melanoma X-actin; Aldo1: aldolase 1, A isoform; Arf5: ADP-ribosylation factor 5; Atf1: activating transcription factor 1; Atf2: activating transcription factor 2; Btf3: basic transcription factor 3a; Bzrp: peripheral-type benzodiazepine receptor; C5r1: complement component 5, receptor 1/G protein-coupled receptor (C5a); Ccnb2: cyclin B2; Cd36l2: CD36 antigen (collagen type I receptor, thrombospondin receptor)-like 2; Cd53: CD53 antigen; Cebpa: CCAAT/enhancer binding protein C/EBP, alpha; Cebpb: CCAAT/enhancer binding protein (C/EBP), beta; Cebpd: CCAAT/enhancer binding protein (C/EBP), delta; Cebpe: CCAAT/enhancer binding protein (C/EBP), epsilon; Cfl1: cofilin 1, nonmuscle; Cmkar4: chemokine (C-X-C) receptor 4; Cmkbr1: chemokine (C-C) receptor 1/Mip1a receptor; Cnlp: cathelin-like protein; Cntf: ciliary neurotropic factor/zinc finger protein PZF; Copa: coatomer protein complex subunit alpha; Cpa3: carboxypeptidase A3, mast cell; Cr2: complement receptor 2; Crhr: corticotropin releasing hormone receptor; Crry: complement receptor–related protein; Csf1r: CSF 1 (M-CSF) receptor/c-fms/CD115; Csf2ra: CSF 2 (GM-CSF) receptor, alpha, low-affinity/CD116; Csf2rb1: CSF 2 (GM-CSF) receptor, beta 2, low-affinity/IL 3 receptor-like protein (AIC2B)/CDw131; Csf2rb2: CSF 2 (GM-CSF) receptor, beta 2, low-affinity/IL-3 receptor (AIC2A); Ctsb: cathepsin B; Ctsc: cathepsin C; Ctsd: cathepsin D; Ctse: cathepsin E; Ctsg: cathepsin G; Ctsh: cathepsin H; Ctsl: cathepsin L; Ctss: cathepsin S; Cybb: cytochrome b-245, beta; Drd2: dopamine receptor 2; E2f1: E2F transcription factor 1; Ear2: eosinophil-associated ribonuclease 2; Ebi3: Epstein-Barr virus–induced gene 3/cytokine receptor–like molecule (EBI3); El2: Balb/c neutrophil elastase; Ela2: elastase 2; Erh: enhancer of rudimentary homolog (Drosophila); Etohi6: ethanol induced 6/sterol regulatory element binding transcription factor 1 (SREBF1) homolog; F2rl2: coagulation factor II (thrombin) receptor–like 2; Fcer1g: Fc receptor, IgE, high affinity I, gamma polypeptide; Fcgr2b: Fc receptor, IgG, low affinity IIb; Fcgr3: Fc receptor, IgG, low affinity III; Fpr1: formyl peptide receptor 1/fMLP receptor; Gabpb1: GA repeat binding protein (GABP-beta1 subunit); Gata2: GATA-binding protein 2; Gnas: guanine nucleotide binding protein, alpha stimulating; Gnb2-rs1: guanine nucleotide binding protein, beta-2, related sequence 1; Gpx3: glutathione peroxidase 3; Grg: related to Drosophila groucho gene; Grid1: glutamate receptor channel subunit delta 1; Grn: granulin; Gstm1: glutathione-S-transferase, mu 1; Gus-s: beta-glucuronidase structural; Gys3: glycogen synthase 3, brain; H2-D: histocompatibility 2, D

region locus 1; Hist2: histone gene complex 2; Hist5–2ax: H2A histone family, member X; Hmgi: high mobility group protein I; Hsp60: heat shock protein, 60 kDa; Htr5a: 5-hydroxytryptamine (serotonin) receptor 5A; Idb1: inhibitor of DNA binding 1/helix-loop-helix DNA binding protein regulator (Id); Idb2: inhibitor of DNA binding 2; Ifngr: interferon gamma receptor; Ifngr2: interferon gamma receptor 2; Ii: Ia-associated invariant chain; Il1a: IL1 alpha; Il1r2: IL1 receptor, type II; Il2rg: IL2 receptor, gamma chain; Il4ra: IL4 receptor, alpha; Il10rb: IL10 receptor, beta; Il17r: IL17 receptor; Irf1: interferon regulatory factor 1; Irf2: interferon regulatory factor-2; Itgb2: integrin beta 2 (Cd18); Itpr5: inositol 1,4,5-trisphosphate receptor (type 2); Jund1: Jun proto-oncogene–related gene d1/transcription factor JUN-D; Klf2: Kruppel-like factor LKLF; L-CCR: lipopolysaccharide inducible C-C chemokine receptor–related; Lcn2: lipocalin 2; Ldlr: low density lipoprotein receptor; Lsp1: Lymphocyte–specific 1/S37/pp52; Lst1: leucocyte–specific transcript 1; Ltb4r: leukotriene B4 receptor; Ltbr: lymphotoxin-beta receptor; Ltf: lactotransferrin; Ly64: lymphocyte antigen 64; Ly6e: lymphocyte antigen 6 complex, locus E; Lyl1: lymphoblastomic leukemia/bHLH factor; Lyzs: lysozyme; M6pr: mannose-6-phosphate receptor, cation dependent; Mad: Max dimerization protein; Man2c1: mannosidase, alpha, class 2C, member 1; Max: Max protein; Maz: MYC-associated zinc finger protein (purine-binding transcription factor); MBP: eosinophil granule major basic protein precursor; Mcpt8: mast cell protease 8; Mll: myeloid/lymphoid or mixed-lineage leukemia; Mmp13: matrix metalloproteinase 13/collagenase; Mmp9: matrix metalloproteinase 9/gelatinase B; Mpo: myeloperoxidase; Myb: myeloblastosis oncogene; Mybl2: myeloblastosis oncogene-like 2; Myc: myelocytomatosis oncogene; Myln: myosin light chain, alkali, nonmuscle; Nfatc2: nuclear factor of activated T cells, cytoplasmic 2; Nfe2: nuclear factor, erythroid-derived 2, 45 kDa; Nfkb1: NF-kappa-B (p105); Ngp: neutrophilic granule protein; NMDRGB: N-methyl-D-aspartate receptor glutamate-binding chain homolog; Npm1: nucleophosmin 1; Nr4a1: nuclear receptor subfamily 4, group A, member 1; Osi: oxidative stress induced; P2rx1: purinergic receptor P2X, ligand-gated ion channel, 1; P2ry2: purinergic receptor P2Y, G-protein–coupled 2; P40–8: P40–8, functional/laminin receptor; Pbx1: pre B-cell leukemia transcription factor 1; Pfc: properdin factor, complement; Pira1: paired-Ig–like receptor A1; Pira5: paired-Ig–like receptor A5; Pira6: paired-Ig–like receptor A6; Pirb: paired-Ig–like receptor B; Plaur: urokinase plasminogen activator receptor; PMI: putative receptor protein (SP: P17152); Pml: promyelocytic leukemia; Prg: proteoglycan, secretory granule; Prg3: proteoglycan 3/eosinophil major basic protein 2; Prtn3: proteinase 3; Psma2: proteasome (prosome, macropain) subunit, alpha type 2; Ptmb4: prothymosin beta 4; Ptprc: protein tyrosine phosphatase, receptor type, C; Rac2: RAS-related C3 botulinum substrate 2; Rarg: retinoic acid receptor, gamma; Rela: avian reticuloendotheliosis viral (v-rel) oncogene homolog A/NF-kappa-B p65; Rpl19: ribosomal protein L19; RPL8: ribosomal protein L8; Rps6ka1: ribosomal protein S6 kinase polypeptide 1; Rps8: ribosomal protein S8; Rtn3: reticulon 3; S100a8: S100 calcium binding protein A8 (calgranulin A); S100a9: S100 calcium-binding protein A9 (calgranulin B); Sdfr2: stromal cell–derived factor receptor 2; Sell: selectin L (lymphocyte adhesion molecule 1); Sema4d: semaphorin 4D; Sepp1: selenoprotein P, plasma, 1; Sfpi1: SFFV proviral integration 1; Shfdg1: split hand/foot deleted gene 1; Slc10a1: solute carrier family 10 (sodium/bile acid cotransporter family), member 1; Slpi: secretory leukocyte protease inhibitor; Sox15: SRY-box containing gene 15; Spi2–1: serine protease inhibitor 2-1; Srb1: scavenger receptor class B1; Stat3: signal transducer and activator of transcription 3; Stat5a: signal transducer and activator of transcription 5A; Stat6: signal transducer and activator of transcription 6; Stra14: basic-helix-loop-helix protein-retinoic acid induced; Tbx1: TBX1 protein/LPS-induced TNF-alpha factor homolog; Tcrgb: T-cell–receptor germline beta-chain gene constant region; Tcrg-V4: T-cell–receptor gamma, variable 4; Tctex1: t-complex testis expressed 1; Tfdp1: transcription factor Dp 1; Tif1b: transcriptional intermediary factor 1, beta; Tlr4: toll-like receptor 4; Tnfrsf1a: TNF receptor superfamily, member 1a; Tnfrsf1b: TNF superfamily, member 1b; Tomm70a: translocase of outer mitochondrial membrane 70 (yeast) homolog A; Tpi: triosephosphate isomerase; Trp53: transformation-related protein 53; Ubb: ubiquitin B; Usf2: upstream transcription factor 2; Ybx1: Y box transcription factor; Ybx3: Y box binding protein; Zfp11–6: zinc finger protein s11–6; Zfp18: zinc finger protein 18 homolog; Zfp36: zinc finger protein 36; Zfp162: zinc finger protein 162; Zfp216: zinc finger protein 216; Zfpm1: zinc finger protein, multitype 1; Znfn1a1: zinc finger protein, subfamily 1A, 1 (Ikaros); Zyx: zyxin.

Opinion

Comparing protein abundance and mRNA expression levels on a genomic scale

Dov Greenbaum*, Christopher Colangelo[†‡], Kenneth Williams[†‡] and Mark Gerstein[†§]

Addresses: *Department of Genetics, [†]Department of Molecular Biophysics and Biochemistry, [‡]HHMI Biopolymer Laboratory and W. M. Keck Foundation Biotechnology Resource Laboratory, and [§]Department of Computer Science, Yale University, New Haven, CT 06520-8114, USA.

Correspondence: Mark Gerstein. E-mail: Mark.Gerstein@yale.edu. Kenneth Williams. E-mail: Kenneth.Williams@yale.edu

Published: 29 August 2003

Genome ***Biology*** 2003, 4:117

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/117

Abstract

Attempts to correlate protein abundance with mRNA expression levels have had variable success. We review the results of these comparisons, focusing on yeast. In the process, we survey experimental techniques for determining protein abundance, principally two-dimensional gel electrophoresis and mass-spectrometry. We also merge many of the available yeast protein-abundance datasets, using the resulting larger 'meta-dataset' to find correlations between protein and mRNA expression, both globally and within smaller categories.

Although some of the underlying technology for quantifying protein abundance was introduced almost thirty years ago [1,2], there has recently been a significant increase in the development of new tools. Concurrently, tools for analyzing mRNA expression are becoming more mainstream. The quantification of both of these molecular populations is not an exercise in redundancy; measurements taken from mRNA and protein levels are complementary and both are necessary for a complete understanding of how the cell works [3]. Additionally, as mRNA is eventually translated into protein, one might assume that there should be some sort of correlation between the level of mRNA and that of protein. Alternatively, there may not be any significant correlation, which, in itself, is an informative conclusion.

The two commonly used high-throughput methods for measuring mRNA expression, microarrays and Affymetrix chips, have both been extensively reviewed elsewhere [4-6]. There are also two basic methods for determining protein abundance; either based on two-dimensional electrophoresis or on mass-spectrometric methods (Table 1). We provide a brief review of these technologies and recent efforts to determine correlations between quantified protein abundances and mRNA expression.

Methods for determining protein levels

Two-dimensional electrophoresis

Determining relative protein expression levels by conventional two-dimensional electrophoresis requires isoelectric focusing, SDS-polyacrylamide gel electrophoresis, staining, fixing, densitometry, and careful matching of the same spots on two or more gels. Differentially expressed spots are then excised and enzymatically digested, and the resulting peptides are identified using mass spectrometry. An attractive aspect of this approach is the low capital equipment cost, but a high level of expertise is needed to obtain reproducible gels, and two-dimensional electrophoresis is generally limited to proteins that are neither too acidic, too basic, nor too hydrophobic, and that are between 10 and 200 kDa in size, so that they are reliably separated on gels. Additionally, this approach detects only those proteins that are expressed at relatively high levels and that have long half-lives [7,8]. In one study using 40 μg yeast lysate, the average protein

Table 1

Overview of selected protein profiling technologies

Technology	Type of labeling required	Ability to detect many post-translational modifications	Biomolecules that are optimally quantified	Approximate dynamic range (and reference)	Number of proteins/spots quantified (and reference)
Two-dimensional gel electrophoresis	Silver staining	Yes	Naturally occurring forms of proteins larger than 10 kDa	10 [9]	1,500 [8]
Differential two-dimensional fluorescence gel electrophoresis (DIGE)	*In vitro* with Cy2, Cy3 or CY5 fluorophores at primary amines	Yes	Naturally occurring forms of proteins larger than 10 kDa	10,000 [9]	1,100 [51]
SELDI- or MALDI-MS disease biomarker discovery	None	Yes	Naturally occurring forms of proteins smaller than 10 kDa	25	Not applicable
Isotope-coded affinity tag (ICAT) - LC/MS	*In vitro* with H^1/D or C^{12}/C^{13} ICAT reagent at cysteine	No	Cysteine-containing tryptic peptides from digests of protein extracts	10,000*	496 [18]
N^{14}/N^{15} - LC/MS	*In vivo* at nitrogens in amino acids	Yes	Tryptic peptides from digests of protein extracts	10,000 [19]	872 [20]

*Assumed to be similar to that for multidimensional protein identification. Abbreviations: SELDI-MS, surface-enhanced laser desorption ionization mass spectrometry; MALDI-MS, matrix-assisted laser desorption ionization mass spectrometry; LC/MS, liquid chromatography and mass spectrometry.

abundance detected was 51,200 copies per cell, with no proteins detected with abundances less than 1,000 copies per cell [8]. Given that 1,500 spots were resolved on a 1.0 pH unit gel [8], several gels covering different pH ranges would be needed to resolve a whole cell lysate. Given these limitations, conventional two-dimensional electrophoresis technology has limited potential for large-scale proteome analysis [8].

Two-dimensional fluorescence-difference gel electrophoresis (DIGE) utilizes mass- and charge-matched, spectrally resolvable fluorescent dyes (such as Cy3 and Cy5) to label two different protein samples *in vitro* prior to two-dimensional electrophoresis. Its main advantage over conventional two-dimensional electrophoresis is that both the control and the experimental sample are run in a single polyacrylamide gel. The samples are then imaged separately but can be perfectly overlaid without any 'warping' of the gels. This substantially raises the confidence with which protein changes between samples can be detected and quantified. Changes in the relative level of expression of a protein may be detected that are as little as 1.2-fold for large-volume spots [9]. Because detection is based on fluorescence, DIGE has a large dynamic range of about 10,000, which permits differential expression analysis of proteins that are present at relatively low copy number [9]. The limit of detection of DIGE for quantifying protein expression ratios is between 0.25 and 0.95 ng protein, which is similar to that for silver staining [9,10]. In a recent study [11], the relative levels of expression of approximately 1,050 protein spots were compared in 250,000 laser-dissected normal versus esophageal carcinoma cells. This analysis identified 58 spots that were up-regulated by more than three-fold and 107 that were down-regulated by more than three-fold in cancer cells.

Mass spectrometric approaches

Disease biomarker discovery

Current approaches to discovering protein or peptide markers of disease involve batch chromatography, matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) and statistical analysis of large numbers of disease versus normal serum or other biological samples. Most recent studies have relied on surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS) [12,13]. The SELDI approach [13] involves using a gold-coated chip with eight or sixteen 2 mm spots that are modified with chromatographic surfaces (for example anionic, cationic, hydrophobic, and so on). After spotting a few microliters of serum, any contaminants and salt are removed by washing with water, and the target is dried by adding a MALDI matrix solution, such as α-cyano-4-hydroxy-cinnamic acid. In a study by Petricoin *et al.* [14] SELDI-MS analysis of serum from 50 control and 50 case samples from patients with ovarian cancer resulted in identifying five peptide biomarkers that ranged in size from 534 to 2,465 Da. The pattern formed by these markers was then used to correctly classify all 50 ovarian cancer samples in a masked set of serum samples from 116 patients who included 50 patients with ovarian cancer and 66 unaffected women. Similar promising results have been reported in studies of serum samples from breast and prostrate cancer patients [12,15]. In a recent study [16], which compared the relative ability of several different statistical approaches to classify samples based on MS data, the disease biomarker approach

was extended to a conventional MALDI-MS platform. Although powerful, the disease biomarker approach does not provide accurate relative amounts of the control versus experimental biomarker, only the relative intensity difference.

Isotope-coded affinity-tag-based protein profiling

While both MALDI-MS-based disease biomarker discovery and DIGE comparatively profile the naturally occurring forms of peptides and proteins, isotope-coded affinity-tag (ICAT) analysis profiles the relative amounts of cysteine-containing peptides derived from tryptic digests of protein extracts. Because only a single tryptic peptide is needed to quantify the expression of the corresponding parent protein, the ICAT reagent utilizes a thiol protein-reactive group that attaches both a biotin tag and either nine ^{12}C (light) or nine ^{13}C (heavy) atoms to each cysteine residue. Following derivatization of the control protein extract with [^{12}C]-ICAT reagent and the experimental extract with [^{13}C]-ICAT reagent, the pooled samples are subjected to trypsin digestion followed by both cation and avidin chromatography. Liquid chromatography and tandem mass spectrometry (LC/MS/MS) is then used to identify ICAT peptide pairs and to quantify the relative $^{12}C/^{13}C$ ratios. It is important to note that the ICAT approach provides the relative expression ratios of individual proteins under two conditions; it does not provide absolute protein concentrations, nor does it provide the ratio of the concentration of one protein relative to another in a single condition. A nice feature of this approach is that the *in vitro* incorporation of a stable isotope into one of the two samples being compared obviates the need to separately analyze the control and experimental samples by MS. Although a tryptic digest of a whole-cell human protein extract might produce more than 500,000 peptides, less than 100,000 of these might be expected to contain cysteine, but based on a search of the SwissProt database [17], less than 5% of human proteins lack cysteine and would therefore be missed (that is, more than 95% of proteins would include at least one cysteine-containing peptide).

ICAT results are analogous to those obtained by the use of two different fluorescent dyes in DNA microarray analysis of mRNA levels or DIGE analysis of protein expression. The largest number of proteins profiled so far using this approach with a single sample are the 491 proteins contained in microsomal fractions of naive and *in vitro* differentiated human myeloid leukemia cells [18].

Multidimensional protein identification technology

Multidimensional protein identification technology (MudPit) is similar to ICAT in that it utilizes cation-exchange prefractionation followed by reverse-phase (RP) high-performance liquid chromatography (HPLC) separation and MS/MS analysis [19]. In contrast to the ICAT approach, however, MudPit analyzes the entire mixture of tryptically digested proteins and utilizes tandemly coupled (cation-exchange followed by reverse-phase) columns. A specific subset of peptides is eluted from the cation-exchange column, using a step gradient of increasing salt concentration, onto the front of the RP column. Peptides are then eluted from the RP column and enter the mass spectrometer for analysis. After the RP gradient is complete, the next step of the salt gradient releases another subset of peptides from the cation-exchange column onto the RP column, and the process repeats itself. Using this approach on the yeast proteome, Wolters *et al.* [19] identified 5,540 unique peptides from 1,484 proteins and demonstrated a dynamic range of detection of 10,000-fold. This method has been extended to comparative protein profiling by using *in vivo* $^{14}N/^{15}N$ metabolic labeling [20,21]. Washburn *et al.* [20] used *Saccharomyces cerevisiae* grown in both ^{14}N- and ^{15}N-containing minimal media, and 2,264 peptides and 872 proteins were uniquely identified. Also, accurate $^{14}N/^{15}N$ quantitation was determined for each peptide with an average standard deviation of 30%.

Comparison of mRNA and protein levels

Even with the significant developments in the technologies used to quantify protein abundance over the past couple of years, protein identification and quantification still lags behind the high-throughput experimental techniques used to determine mRNA expression levels. Yet, while mRNA expression values have shown their usefulness in a broad range of applications, including the diagnosis and classification of cancers [22,23], these results are almost certainly only correlative, rather than causative; in the end it is most probably the concentration of proteins and their interactions that are the true causative forces in the cell, and it is the corresponding protein quantities that we ought to be studying.

Primarily because of a limited ability to measure protein abundances, researchers have tried to find correlations between mRNA and the limited protein expression data, in the hope that they could determine protein abundance levels from the more copious and technically easier mRNA experiments. Alternatively, if there is definitively no correlation between mRNA and protein data, both quantities could be used as independent sources of information for use in machine-learning algorithms, for example, to predict protein interactions. To date, there have been only a handful of efforts to find correlations between mRNA and protein expression levels, most notably in human cancers and yeast cells; for the most part, they have reported only minimal and/or limited correlations.

One of the earliest analyses of correlation looked at 19 proteins in the human liver. Anderson and Seilhamer [24] found a somewhat positive correlation of 0.48. Another limited analysis, of the three genes *MMP-2*, *MMP-9* and *TIMP-1* in human prostate cancers, showed no significant relationship [25]. An additional cancer study [26] showed a

significant correlation in only a small subset of the proteins studied. Conversely, Ornstoft *et al.* [27] found highly significant correlations in human carcinomas when looking at changes in mRNA and protein expression levels.

Protein and mRNA correlations in yeast

Many of the present efforts at correlating mRNA and protein expression have been conducted in yeast using two-dimensional electrophoresis techniques. In particular, Gygi *et al.* [7] found that even similar mRNA expression levels could be accompanied by a wide range (up to 20-fold difference) of protein abundance levels, and *vice versa*. These results contrast with those of Futcher *et al.* [28], who found relatively high correlations (r = 0.76) after transforming the data to normal distributions. In a previous analysis [29], we merged the data from both of these datasets (referred to as 2DE-1 [7] and 2DE-2 [28]), comparing the resulting new larger protein abundance set ('merged data-set 1') with a comprehensive mRNA expression dataset. The mRNA expression reference set was constructed through iteratively combining, in a non-trivial fashion, three sets that used Affymetrix chips and a SAGE dataset [29]. Using these reference datasets, we were able to do an all-against-all comparison of mRNA and protein expression levels, in addition to a number of analyses comparing protein and mRNA expression using smaller, but broad categories [29,30].

Given the difficult, laborious, and limiting nature of two-dimensional electrophoresis analysis, many of the newer protein abundance determinations have been done using MudPit and derivative technologies. Washburn *et al.* [31] used MudPit to analyze and detect 1,484 arbitrary proteins: they were able to detect a somewhat random sampling of proteins independent of abundance, localization, size or hydrophobicity (we refer to this dataset as MudPit-1). In a further experiment, the authors, comparing expression ratios for both proteins and mRNA levels, found that although they could not find correlations for individual loci, they could find overall correlations when looking at pathways and complexes of proteins that functioned together [21]. Peng *et al.* [32] analyzed 1,504 yeast proteins with a false-positive rate - misidentification of a protein - of less than 1% (we refer to this dataset as MudPit-2). In their analysis [32], they contrasted their methodology with that of Washburn *et al.* [31] with which there was significant overlap of proteins.

A new merged dataset

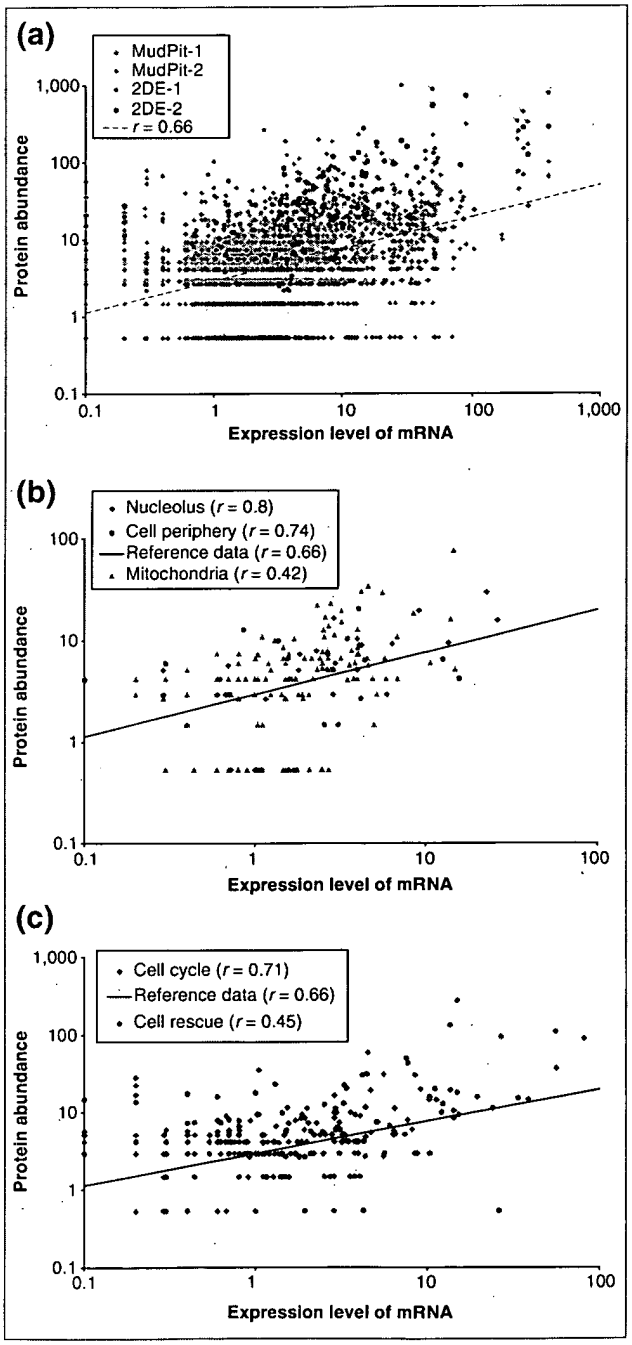
Expanding upon our previous merged dataset, we constructed a new merged dataset (merged data set-2) using the two two-dimensional electrophoresis and two MudPit datasets described above. Succinctly (more information is available on our website at [33]), we transformed each of the protein-abundance datasets into more quantitative data by fitting each protein dataset individually onto the reference mRNA expression dataset. The MudPit-1 dataset was also fitted onto the more finely grained MudPit-2 dataset. Each of the new, fitted datasets was then inversely transformed back into protein space. These derived protein datasets were then combined into a larger reference dataset; when we had more than one abundance value for an open reading frame (ORF), we chose the value from the dataset according to a prescribed quality ranking (see Figure 1). The resulting set contained protein abundance information for approximately 2,000 ORFs. (One caveat with the MudPit data: while quantitative analysis can be subsequently done on the results of MudPit experiments, MudPit data alone are only semi-quantitative, in that the number of peptides determined is relative to the actual protein abundance within the cell [31]. Some may therefore argue that MudPit alone is not optimal for a comparison with mRNA data. Nevertheless, we feel that our methodical merging process creates a quantitative and representative dataset that can be compared with the mRNA expression data.) Using the resulting data we could compare mRNA expression and protein abundance globally (Figure 1a) as well as looking at smaller, broad categories, such as function or localization (see Figure 1b,c). In particular, we show that some localization categories - for example, the nucleolus - have significantly higher correlations than the global correlation. Other localizations may present less of a correlation between mRNA and protein data - for example, the mitochondria - possibly reflecting the heterogeneous nature and function of the latter organelle. In terms of MIPS functional categories [34,35], we show that although some categories, such as cell rescue, show a lower correlation than the whole merged set, other functional categories, such as cell cycle, show a significant increase in correlation. Logically, this increased correlation reflects the co-regulated nature of the proteins in this functional category.

Reasons for the absence of correlation

There are presumably at least three reasons for the poor correlations generally reported in the literature between the level of mRNA and the level of protein, and these may not be mutually exclusive. First, there are many complicated and varied post-transcriptional mechanisms involved in turning mRNA into protein that are not yet sufficiently well defined to be able to compute protein concentrations from mRNA; second, proteins may differ substantially in their *in vivo* half lives; and/or third, there is a significant amount of error and noise in both protein and mRNA experiments that limit our ability to get a clear picture [36,37].

Examining the first option - that there are a number of complex steps between transcription and translation - we looked at correlations between mRNA and protein abundance for those ORFs that had varied or steady levels of mRNA expression over the course of the cell cycle [38]. To normalize for the varied degrees of expression for different ORFs, we took the standard deviation divided by the average expression level as representative of the variation of each ORF over the course of the yeast cell cycle (Figure 2). Broadly speaking, the cell can control the levels of protein at

the transcriptional level and/or at the translational level. Logically, we would assume that those ORFs that show a large degree of variation in their expression are controlled at the transcriptional level - the variability of the mRNA expression is indicative of the cell controlling mRNA expression at different points of the cell cycle to achieve the resulting and desired protein levels. Thus we would expect, and we found, a high degree of correlation (r = 0.89) between the reference mRNA and protein levels for these particular ORFs; the cell has already put significant energy into dictating the final level of protein through tightly controlling the mRNA expression, and we assume that there would then be minimal control at the protein level. In contrast, those genes that show minimal variation in their mRNA expression throughout the cell cycle are more likely to have little or no correlation with the final protein level; the cell would be controlling these ORFs at the translational and/or post-translational level, with the mRNA levels being somewhat independent of the final protein concentration. And indeed, we found only minimal correlation between protein and mRNA expression for these ORFs (r = 0.2).



Furthermore, we found that those ORFs that have higher than average levels of ribosomal occupancy - that is that a large percentage of their cellular mRNA concentration is associated with ribosomes (being translated) - have well correlated mRNA and protein expression levels (Figure 2). These cases probably represent a situation wherein the cell, having significantly controlled the mRNA expression to produce a specific level of protein, will probably not also employ mechanisms to control the translation. Alternatively, those proteins that have very low occupancy rates have uncorrelated mRNA and protein expression; thus, given that the cell has not tightly controlled the mRNA expression for this ORF, it will dictate the resulting protein levels through rigorous controls of its translation (that is, through tight limits on occupancy) [39].

A second option for a general lack of correlation between mRNA and protein abundance may be that proteins have very different half-lives as the result of varied protein synthesis and degradation. Protein turnover can vary significantly depending on a number of different conditions [40]; the cell can control

Figure 1
Comparison of mRNA expression and protein abundance. **(a)** A plot comparing our mRNA reference expression set [29] with our newly compiled protein abundance dataset. The mRNA axis is in copies per cell; the protein axis is in thousand copies per cell. The protein dataset is the result of iteratively fitting two MudPit datasets (MudPit-1 [32] and MudPit-2 [31]) and two two-dimensional electrophoresis datasets (2DE-1 [7] and 2DE-2 [28]). Given the semi-quantitative nature of the MudPit data [31], we transformed the data into a more quantitative set by fitting each set individually onto our reference mRNA expression dataset. In addition, we fit the MudPit-1 dataset onto the more finely-grained MudPit-2 dataset. Each of the datasets was then moved back into 'protein space' using an inverse transformation derived from the 2DE-1 set, as this set has the most precise values. These datasets were then combined into the new reference abundance dataset. In cases in which there were overlapping values for a given ORF we used the dataset in accord with the following ordering: 2DE-1, 2DE-2, MudPit-2, MudPit-1. The resulting reference protein abundance dataset (N = 2044) had a correlation of 0.66 with the mRNA reference dataset. **(b,c)** Additionally, we show that when looking at specific subsets (subcellular localization [52] or functional groups [34,35]) we can find both higher and lower correlations amongst these groups. The lower correlations are generally reflective of a more heterogeneous category. This analysis indicates that while correlations may be weak when looking at the global data, we tend to find higher correlations when looking at smaller well-defined subsets of ORFs. Further analysis is available at [33].

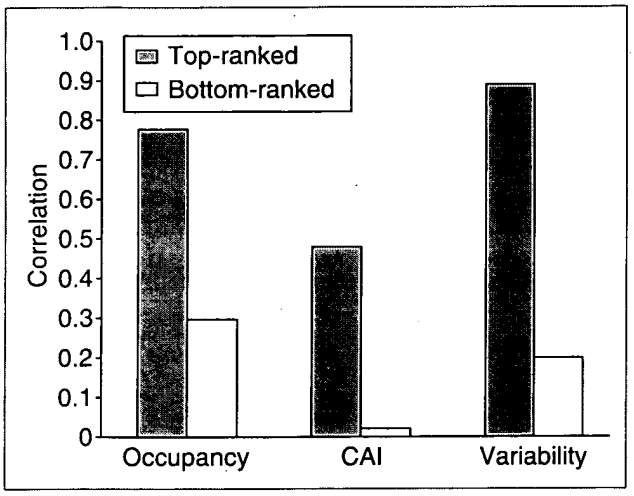


Figure 2
The differences in correlation between mRNA and protein expression values using novel categories. We see significant differences when looking at the highest and lowest ranking of groups of ORFs in the following categories: occupancy, CAI (codon adaptation index) value [45-47] and variability. Occupancy refers to the percentage of transcripts associated with ribosomes; we compared the correlation between the top 100 ORFs and the bottom 100 in terms of occupancy (r =0.78 versus 0.30). For the CAI, we compared the correlation between mRNA and protein for those ORFs with the highest CAI and those with the lowest (r = 0.48 versus 0.02). Variability refers to the normalized standard deviation (that is, the standard deviation divided by the average expression level) for all ORFs in the cell-cycle expression dataset of Cho *et al.* [38]. Here, we compared the correlations between protein abundance and mRNA expression for the most variable compared with the least variable proteins (r = 0.89 versus 0.20). We found significant differences between the correlations of mRNA and protein levels for the top and bottom ranking populations for each of the comparisons.

the rates of degradation or synthesis for a given protein, and there is significant heterogeneity even within proteins that have similar functions [41]. Recent efforts have been made to computationally measure these rates [42].

Simplistically, it can be presumed that the change in a protein's concentration over time will be equal to the rate of translation minus the rate of degradation. By analogy to concepts in chemical kinetics, we can approximate this equation: $dP(i,t)/dt = SE(i,t) - DP(i,t)$, where P is protein abundance i at time t, E is the mRNA expression level of protein P, S is a general rate of protein synthesis per mRNA, and D is a general rate of protein degradation per protein [43]. Additionally there are some experimental methods that can also be used to measure turnover and the translational control of protein levels [41-44].

Given the degenerate nature of the genetic code, there are many synonymous codons (codons that translate into the same amino acid). As the cell is biased in its usage of synonymous codons - that is, the usage of a subset of codons results in a higher level of mRNA expression, possibly as a result of differing cellular tRNA levels [45] - the codon adaptation index (CAI), a measurement of codon usage, can be used to predict the expression of a gene [46] (we recently calculated new parameters for this model, with some improvement in predictive strength [47]). It is thought that the CAI will correlate differently with mRNA levels than with protein abundance levels due, in part, to protein turnover rates [48]. Ranking the ORFs in terms of their CAI value, we found that although those ORFs that ranked the highest in terms of CAI did not show a very strong correlation between mRNA and protein levels, they nevertheless showed a significantly higher correlation than ORFs that were ranked as having the lower CAI values (r = 0.48 versus 0.02). The low correlations reflect the fact that the CAI will correlate differently for protein and mRNA values because of the additional cellular controls on protein translation, namely the effect of protein turnover rates. Nevertheless, the sizable difference in correlations between the two groups of ORFs with high- and low-ranking CAI values (Figure 2) shows that there is some relationship between mRNA and protein values, possibly indicating that highly expressed genes tend to result in a more correlated level of protein abundance than lower expressed ones.

Correlations have been found between the mRNA expression levels of different protein subunits within protein complexes [49]. This implies that there should be, in general, a correlation between mRNA and protein abundance, as these subunits provide a special case as they have to be available in stoichiometric amounts of proteins for the complexes to function. Thus, we believe that a major limitation to finding correlations is the degree of natural and manufactured systematic noise in mRNA and protein expression experiments. There is a continued effort to both describe and reduce this noise [50]. Meanwhile, in an attempt to get around the noise one could look at broad categories of proteins - for example, groups defined by function, structure, or localization - such that the background noise is cancelled out to some degree [29].

Although proteomics is still in its infancy, given the pace of technological advancement in protein quantification, mRNA expression analysis and noise reduction, more comprehensive correlation studies will soon be feasible. This will allow for more robust analyses of the relationship between mRNA expression and protein abundance values. Finally, to be fully able to understand the relationship between mRNA and protein abundances, the dynamic processes involved in protein synthesis and degradation have to be better understood; is the protein level changing because of a change in the rate of protein synthesis, or mRNA, or protein turnover? These questions need to be looked into further before we can appreciate in full the relationship between mRNA and protein abundance levels.

Acknowledgements

This project was funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28186.

References

1. O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins.** *J Biol Chem* 1975, **250**:4007-4021.
2. Klose J: **Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals.** *Humangenetik* 1975, **26**:231-243.
3. Hatzimanikatis V, Choe LH, Lee KH: **Proteomics: theoretical and experimental considerations.** *Biotechnol Prog* 1999, **15**:312-318.
4. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: **Microarrays: biotechnology's discovery platform for functional genomics.** *Trends Biotechnol* 1998, **16**:301-306.
5. McGall GH, Christians FC: **High-density genechip oligonucleotide probe arrays.** *Adv Biochem Eng Biotechnol* 2002, **77**:21-42.
6. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21**:33-37.
7. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.
8. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R: **Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.** *Proc Natl Acad Sci USA* 2000, **97**:9390-9395.
9. Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, Pognan F, Hawkins E, Currie I, Davison M: **Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology.** *Proteomics* 2001, **1**:377-396.
10. Gharbi S, Gaffney P, Yang A, Zvelebil MJ, Cramer R, Waterfield MD, Timms JF: **Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system.** *Mol Cell Proteomics* 2002, **1**:91-98.
11. Zhou G, Li H, DeCamp D, Chen S, Shu H, Gong Y, Flaig M, Gillespie JW, Hu N, Taylor PR, *et al.*: **2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers.** *Mol Cell Proteomics* 2002, **1**:117-124.
12. Adam BL, Vlahou A, Semmes OJ, Wright GL Jr.: **Proteomic approaches to biomarker discovery in prostate and bladder cancers.** *Proteomics* 2001, **1**:1264-1270.
13. Issaq HJ, Veenstra TD, Conrads TP, Felschow D: **The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification.** *Biochem Biophys Res Commun* 2002, **292**:587-592.
14. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, *et al.*: **Use of proteomic patterns in serum to identify ovarian cancer.** *Lancet* 2002, **359**:572-577.
15. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer.** *Clin Chem* 2002, **48**:1296-1304.
16. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data.** *Bioinformatics*, in press.
17. **SwissProt** [http://www.expasy.ch/sprot/]
18. Han DK, Eng J, Zhou H, Aebersold R: **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry.** *Nat Biotechnol* 2001, **19**:946-951.
19. Wolters, DA Washburn MP, Yates JR 3rd: **An automated multidimensional protein identification technology for shotgun proteomics.** *Anal Chem* 2001, **73**:5683-5690.
20. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR 3rd: **Analysis of quantitative proteomic data generated via multidimensional protein identification technology.** *Anal Chem* 2002, **74**:1650-1657.
21. Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR 3rd: **Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2003, **100**:3107-3112.
22. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
23. Macgregor PF, Squire JA: **Application of microarrays to the analysis of gene expression in cancer.** *Clin Chem* 2002, **48**:1170-1177.
24. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in human liver.** *Electrophoresis* 1997, **18**:533-537.
25. Lichtinghagen R, Musholt PB, Lein M, Romer A, Rudolph B, Kristiansen G, Hauptmann S, Schnorr D, Loening SA, Jung K: **Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue.** *Eur Urol* 2002, **42**:398-406.
26. Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, *et al.*: **Discordant protein and mRNA expression in lung adenocarcinomas.** *Mol Cell Proteomics* 2002, **1**:304-313.
27. Orntoft TF, Thykjaer T, Waldman FM, Wolf H, Celis JE: **Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas.** *Mol Cell Proteomics* 2002, **1**:37-45.
28. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**:7357-7368.
29. Greenbaum D, Jansen R, Gerstein M: **Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts.** *Bioinformatics* 2002, **18**:585-596.
30. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M: **Interrelating different types of genomic data, from proteome to secretome: 'oming in on function.** *Genome Res* 2001, **11**:1463-1468.
31. Washburn MP, Wolters D, Yates JR 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19**:242-247.
32. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP: **Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome.** *J Proteome Res* 2003, **2**:43-50.
33. **Gerstein Lab - Supplementary data tables** [http://bioinfo.mbb.yale.edu/expression/mrna-v-protein/]
34. **MIPS database** [http://mips.gsf.de/]
35. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30**:31-34.
36. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17**:509-519.
37. Szallasi Z: **Genetic network analysis in light of massively parallel biological data acquisition.** *Pac Symp Biocomput* 1999, 5-16.
38. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.
39. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D: **Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2003, **100**:3889-3894.
40. Glickman MH, Ciechanover A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction.** *Physiol Rev* 2002, **82**:373-428.
41. Pratt JM, Petty J, Riba-Garcia I, Robertson DH, Gaskell SJ, Oliver SG, Beynon RJ: **Dynamics of protein turnover, a missing dimension in proteomics.** *Mol Cell Proteomics* 2002, **1**:579-591.
42. Lian Z, Kluger Y, Greenbaum DS, Tuck D, Gerstein M, Berliner N, Weissman SM, Newburger PE: **Genomic and proteomic analysis of the myeloid differentiation program: global analysis of gene expression during induced differentiation in the MPRO cell line.** *Blood* 2002, **100**:3209-3220.
43. Gerner C, Vejda S, Gelbmann D, Bayer E, Gotzmann J, Schulte-Hermann R, Mikulits W: **Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling.** *Mol Cell Proteomics* 2002, **1**:528-537.
44. Serikawa KA, Xu XL, MacKay VL, Law GL, Zong Q, Zhao LP, Bumgarner R, Morris DR: **The transcriptome and its translation during recovery from cell cycle arrest in *Saccharomyces cerevisiae*.** *Mol Cell Proteomics* 2003, **2**:191-204.
45. Bennetzen JL, Hall BD: **Codon selection in yeast.** *J Biol Chem* 1982, **257**:3026-3031.
46. Sharp PM, Li WH: **The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**:1281-1295.

47. Jansen R, Bussemaker HJ, Gerstein M: **Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models.** *Nucleic Acids Res* 2003, **31**:2242-2251.
48. Coghlan A, Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, **16**:1131-1145.
49. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res* 2002, **12**:37-46.
50. Qian J, Kluger Y, Yu H, Gerstein M: **Identification and correction of spurious spatial correlations in microarray data.** *Biotechniques* 2003, 35:42-44.
51. Yan, JX, Devenish, AT, Wait, R, Stone, T, Lewis, S, Fowler, S: **Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*.** *Proteomics* 2002, **2**:1682-98.
52. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, *et al.*: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16**:707-719.

Discordant Protein and mRNA Expression in Lung Adenocarcinomas*

Guoan Chen‡, Tarek G. Gharib‡, Chiang-Ching Huang§, Jeremy M. G. Taylor§, David E. Misek¶, Sharon L. R. Kardia‖, Thomas J. Giordano**, Mark D. Iannettoni‡, Mark B. Orringer‡, Samir M. Hanash¶, and David G. Beer‡ ‡‡

The relationship between gene expression measured at the mRNA level and the corresponding protein level is not well characterized in human cancer. In this study, we compared mRNA and protein expression for a cohort of genes in the same lung adenocarcinomas. The abundance of 165 protein spots representing 98 individual genes was analyzed in 76 lung adenocarcinomas and nine non-neoplastic lung tissues using two-dimensional polyacrylamide gel electrophoresis. Specific polypeptides were identified using matrix-assisted laser desorption/ionization mass spectrometry. For the same 85 samples, mRNA levels were determined using oligonucleotide microarrays, allowing a comparative analysis of mRNA and protein expression among the 165 protein spots. Twenty-eight of the 165 protein spots (17%) or 21 of 98 genes (21.4%) had a statistically significant correlation between protein and mRNA expression ($r > 0.2445$; $p < 0.05$); however, among all 165 proteins the correlation coefficient values (r) ranged from −0.467 to 0.442. Correlation coefficient values were not related to protein abundance. Further, no significant correlation between mRNA and protein expression was found ($r = -0.025$) if the average levels of mRNA or protein among all samples were applied across the 165 protein spots (98 genes). The mRNA/protein correlation coefficient also varied among proteins with multiple isoforms, indicating potentially separate isoform-specific mechanisms for the regulation of protein abundance. Among the 21 genes with a significant correlation between mRNA and protein, five genes differed significantly between stage I and stage III lung adenocarcinomas. Using a quantitative analysis of mRNA and protein expression within the same lung adenocarcinomas, we showed that only a subset of the proteins exhibited a significant correlation with mRNA abundance. ***Molecular & Cellular Proteomics 1:304–313, 2002.***

Lung cancer is the leading cause of cancer death for both men and women in the United States. Adenocarcinomas of the lung comprise ~40% of all new cases of non-small cell lung cancer and are now the most common histologic type. Functional genomics, broadly defined as the comprehensive analysis of genes and their products, have become a recent focus of the life sciences (1). Application of these approaches to lung adenocarcinomas has the potential to aid in the identification of high risk patients with resectable early stage lung cancer that may benefit from adjuvant therapy, as well as to identify new therapeutic targets. In human lung cancer, however, little is currently understood regarding the relationship between gene expression as determined by measuring mRNA levels and the corresponding abundance of the protein products.

A number of powerful techniques for analysis of gene expression have been used including differential display (2), serial analysis of gene expression (3), DNA microarrays (4), and proteomics via two-dimensional polyacrylamide gel electrophoresis and mass spectrometry (5). Bioinformatics tools have also been developed to help determine quantitative mRNA/protein expression profiles of all types of cells and tissues (6) and now can be applied to benign and malignant tumors. DNA microarrays (cDNA and oligonucleotide) permit the parallel assessment of thousands of genes and have been utilized in gene expression monitoring (7), polymorphism analysis (8), and DNA sequencing (9). Recent studies have focused on classification or identification of subgroups of lung tumors using DNA microarrays (10, 11). The use of mRNA expression patterns by themselves, however, is insufficient for understanding the expression of protein products, as additional post-transcriptional mechanisms, including protein translation, post-translational modification, and degradation, may influence the level of a protein present in a given cell or tissue. Proteomic analyses, a complementary technology to DNA microarrays for monitoring gene expression, involves protein separation and quantitative assessment of protein spots using 2D[1]-PAGE and protein identification using mass spectrometry. By combining proteomic and transcriptional analyses of the same samples, however, it may be possible to understand the complex mechanisms influencing protein expression in human cancer.

In this study, we determined mRNA and protein levels for 165 proteins (98 genes) in 76 lung adenocarcinomas and nine

From the Departments of ‡Surgery, §Biostatistics, ‖Epidemiology, **Pathology, and ¶Pediatrics, University of Michigan, Ann Arbor, Michigan 48109

Received, January 21, 2002, and in revised form, March 4, 2002
Published, MCP Papers in Press, March 12, 2001, DOI 10.1074/mcp.M200008-MCP200

[1] The abbreviations used are: 2D, two-dimensional; MALDI-MS, matrix-assisted laser desorption/ionization mass spectrometry.

TABLE I

Correlation coefficients of protein and mRNA where only one spot was present on 2D gels

r*, correlation coefficient value > 0.2445; $p < 0.05$. Values in boldface are significant at $p < 0.05$.

Spot	Unigene	Gene name	r*	Protein name
1104	Hs.184510	SFN	**0.4337**	14-3-3 σ
0994	Hs.77840	ANXA4	**0.4219**	Annexin IV
1314	Hs.10958	DJ-1	**0.3982**	DJ-1 protein/MER5
1454	Hs.75428	SOD1	**0.3863**	Superoxide dismutase (Cu-Zn)
1638	Hs.227751	LGALS1	**0.3318**	Galectin 1
0264	Hs.129548	HNRPK	**0.3034**	Transformation up-regulated nuclear protein
1405	Hs.111334	FTL	**0.2849**	Ferritin light chain
0963	Hs.300711	ANXA5	**0.2468**	Annexin V
1252	Hs.4745	PSMC	**0.2445**	26 S proteasome p28
0906	Hs.234489	LDHB	0.4420	L-lactate dehydrogenase H chain (LDH-B)
1171	Hs.241515	COX11	0.2310	COX 11
1160	Hs.181013	PGAM1	0.2023	Phosphoglycerate mutase
0759	Hs.74635	DLD	0.1965	Dihydrolipoamide dehydrogenase precursor
1193	Hs.83383	AOE372	0.1932	Antioxidant enzyme AOE372
0172	Hs.3069	HSPA9B	0.1872	GRP75
0777	Hs.979	PDHB	0.1855	Pyruvate dehydrogenase E1-β subunit precursor
1249	Hs.226795	GSTP1	0.1773	Glutathione *S*-transferase pi (GST-pi)
1685	Hs.76136	TXN	0.1732	Thioredoxin
1205	Hs.82314	HPRT1	0.1588	HG phosphoribosyltransferase
1230	Hs.279860	TPT1	0.1466	Translationally controlled tumor protein (TCTP)
0603	Hs.181357	LAMR1	0.1463	LAMR
1358	Hs.28914	APRT	0.1399	Adenine phosphoribosyl transferase
1410	Hs.82113	DUT	0.1213	dUTP pyrophosphatase (dUTPase)
1825	Hs.112378	LIMS1	0.1213	Pinch-2 protein
0871	Hs.250502	CA8	0.1122	Carbonic anhydrase-related protein; Syntaxin
0289	Hs.82916	CCT6A	0.1106	Chaperonin-like protein
1143	Hs.11465	GSTTLp28	0.0997	Glutathione *S*-transferase homolog (GST homolog)
1456	Hs.118638	NME1	0.0932	Nm23 (NDPKA)
1598	Hs.278503	RIG	0.0905	RIIG (U32331)
1354	Hs.89761	ATP5D	0.0904	FIFO-type ATP synthase subunit d
1445	Hs.155485	HIP2	0.0843	Huntingtin interacting protein 2 (HIP2)
1479	Hs.177486	APP	0.0746	Amyloid B4A
0608	Hs.182265	KRT19	0.0439	Cytokeratin 19
1071	Hs.10842	RAN	0.0277	GTP-binding nuclear protein RAN(TC4)
0991	Hs.297939	CTSB	0.0254	Cathepsin B
0842	Hs.77274	PLAU	0.0248	Urokinase plasminogen activator
0823	Hs.198248	B4GALT1	0.0183	β 1,4-galactosyl transferase
0613	Hs.1247	APOA4	0.0176	Apolipoprotein A4 (ApoA4)
1338	Hs.104143	CLTA	0.0123	Clathrin light chain A
0902	Hs.5123	SID6-306	0.0117	Cytosolic inorganic pyrophosphatase
1688	Hs.1473	GRP	−0.0040	Preprogastrin-releasing peptide
0265	Hs.274402	HSPA1B	−0.0071	Heat shock-induced protein
1414	Hs.77541	ARF5	−0.0096	ADP-ribosylation factor 1
0710	Hs.97206	HIP1	−0.0114	Huntingtin interacting protein 1 (HIP1)
0532	Hs.170328	MSN	−0.0132	Moesin/E
0525	Hs.284255	ALPP	−0.0148	Alkaline phosphate, placental
0513	Hs.76901	PDIR	−0.0289	Protein disulfide isomerase-related protein 5
1659	Hs.256697	HINT	−0.0312	Protein kinase C inhibitor
1262	Hs.7016	RAB7	−0.0362	Rab 7 protein
0190	Hs.184411	ALB	−0.0470	Albumin
0948	Hs.2795	LDHA	−0.0549	Lactate dehydrogenase-A (LDHA)
0502	Hs.180532	GPI	−0.0575	Hsp89
0152	Hs.75410	HSPA5	−0.0640	GRP78
1054	Hs.74276	CLIC1	−0.0686	Nuclear chloride channel (RNCC protein)
0709	Hs.253495	SFTPD	−0.0936	Pulmonary surfactant protein D
0867	Hs.78996	PCNA	−0.0982	PCNA
0165	Hs.180414	HSPA8	−0.1014	Heat shock cognate protein, 71 kDa
1109	Hs.75103	YWHAZ	−0.1018	14-3-3 ζ/Δ
0137	Hs.554	SSA2	−0.1032	Ro/ss-A antigen

TABLE I—*continued*

Spot	Unigene	Gene name	r*	Protein name
0278	Hs.4112	TCP1	−0.1237	T-complex protein I, α subunit
1769	Hs.9614	NPM1	−0.1738	B23/numatrin
0089	Hs.74335	HSPCB	−0.2049	Hsp90
2511	Hs.153179	FABP5	−0.2109	E-FABP/FABP5
1739	Hs.16488	CALR	−0.2344	Calreticulin 32
1138	Hs.301961	GSTM4	−0.2438	Glutathione *S*-transferase M4 (GST m4)
2533	Hs.77060	PSMB6	−0.2512	Macropain subunit Δ

non-neoplastic lung tissues. Protein levels were determined using quantitative 2D-PAGE analysis, and the separated protein polypeptides were identified using matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). The corresponding mRNA levels for the identified proteins within the same samples were determined using oligonucleotide microarrays. Correlation analyses showed that protein abundance is likely a reflection of the transcription for a subset of proteins, but translation and post-translational modifications also appear to influence the expression levels of many individual proteins in lung adenocarcinomas.

EXPERIMENTAL PROCEDURES

Tissues—Fifty-seven stage I and 19 stage III lung adenocarcinomas, as well as nine non-neoplastic lung tissue samples, were used for protein and mRNA analyses. Patient consent was obtained, and the project was approved by the Institutional Review Board. All tissues were obtained after resection at the University of Michigan Health System between May 1991 and July 1998. Tissues were all snap-frozen in liquid nitrogen and then stored at −80 °C. The patients included 46 females and 30 males ranging in age from 40.9 to 84.6 (average 63.8) years. Most patients (66/76) demonstrated a positive smoking history. Sixty-one tumor samples were classified as bronchial-derived, 14 were classified as bronchoalveolar, and one had both features. Eighteen tumor samples were classified as well differentiated, 38 were classified as moderate, and 19 were classified as poorly differentiated adenocarcinomas. Hematoxylin-stained cryostat sections (5 μm), prepared from the same tumor pieces to be utilized for protein and mRNA isolation, were evaluated by a pathologist and compared with hematoxylin- and eosin-stained sections made from paraffin blocks of the same tumors. Specimens were excluded from analysis if they showed unclear or mixed histology (*e.g.* adenosquamous), tumor cellularity less than 70%, potential metastatic origin as indicated by previous tumor history, extensive lymphocytic infiltration, or fibrosis or if the patient had received prior chemotherapy or radiotherapy.

Oligonucleotide Array Hybridization—The HuGeneFL oligonucleotide arrays (Affymetrix, Santa Clara, CA) containing 6800 genes were used in this study. Total RNA was isolated from all samples using Trizol reagent (Invitrogen). The resulting RNA was then subjected to further purification using RNeasy spin columns (Qiagen). Preparation of cRNA, hybridization, and scanning of the HuGeneFL arrays were performed according to the manufacturer's protocol (Affymetrix, Santa Clara, CA). Data analysis was performed using GeneChip 4.0 software. The gene expression profile of each tumor was normalized to the median gene expression profile for the entire sample. Details of data trimming and normalization are described elsewhere (11).

2D-PAGE and Quantitative Protein Analysis—Tissue for both protein and mRNA isolation came from contiguous areas of each sample. Protein separation using 2D-PAGE, silver staining, and digitization were performed as described previously (12, 13). Our 2D-PAGE system allows us to run 20 gels at one time (one batch). Spot detection and quantification were accomplished utilizing Bio Image Visage System software (Bioimage Corp., Ann Arbor, MI). The integrated intensity of each spot was calculated as the measured optical density units × mm^2. Of the total possible 2000 spots detectable on each gel, 820 spots on the gel of each sample were matched using a Gel-ed match program with the same spots on a chosen "master" gel. In each sample, 250 ubiquitously expressed reference spots were used to adjust for variations between gels, such as that created by subtle differences in protein loading or gel staining. Slight differences because of batch were corrected after spot-size quantification.

Mass Spectrometry and 2D Western Blotting—Preparative 2D gels were run using extracts from A549 lung adenocarcinoma cells (obtained from ATCC) and using the identical experimental conditions as the analytical 2D gels, except 30% more protein was loaded. The resolved protein gels were silver-stained using successive incubations in 0.02% sodium thiosulfate for 2 min, 0.1% silver nitrate for 40 min, and 0.014% formaldehyde plus 2% sodium carbonate for 10 min. For protein identification, protein polypeptides underwent trypsin digestion followed by MALDI-MS using a MALDI-TOF Voyager-DE mass spectrometer (Perseptive Biosystems, Framingham, MA). The masses were compared with known trypsin digest databases using the MS-FIT database (University of California, San Francisco; prospector.ucsf.edu/ucsfhtml3.2/msfit.htm). Some of the polypeptides included in the analysis had been identified prior to this study on the basis of sequencing (14). The identified protein spots used in this paper are shown in Fig. 1*A*. The method for 2D-PAGE Western blot verification was as described previously (15). The 2D Western blots of GRP58 and Op18 are shown in Fig. 1, *C* and *E*; the others, such as GRP78, GRP75, HSP70, HSC70, KRT8, KRT18, KRT19, Vimentin, ApoJ, 14-3-3, Annexin I, Annexin II, PGP9.5, DJ-1, GST-pi, and PGAM, are described elsewhere.[2]

Statistical Analysis—Missing values were replaced with the mean value of the protein spot. The transform $x \rightarrow \log(1 + x)$ was applied to normalize all protein expression values. The relationship between protein and mRNA expression levels within the same samples was examined using the Spearman correlation coefficient analysis (16). To identify potentially significant correlations between gene and protein expression, we used an analytical strategy similar to SAM (significance analysis of microarrays) (17), which uses a permutation technique to determine the significance of changes in gene expression between different biological states. To obtain permuted correlation coefficients between gene and protein expression, genes were exchanged first in such a way that permutated correlation coefficient were calculated based on pseudo pairs of genes and proteins. The distribution of permutated correlation coefficients became stable after 60 permutations. This procedure was then repeated 60 times to obtain 60 sets of permutated correlation coefficients. For each of the 60 permutations, the correlations of genes and proteins were ranked

[2] Chen *et al.*, submitted for publication.

TABLE II

Correlation coefficients of protein and mRNA where multiple isoforms were present on 2D gels

r^*, correlation coefficient value > 0.2445; $p < 0.05$. Values in boldface are significant at $p < 0.05$.

Spot	Unigene	Gene name	r^*	Protein name
1494	Hs.81915	LAP18	**0.4003**	OP18 (Stathmin)
0957	Hs.77899	TPM1	**0.3930**	Tropomyosins 1-5
0353	Hs.289101	GRP58	**0.3802**	Protease disulfide isomerase (GRP58)
0855	Hs.169476	GAPD	**0.3693**	Glyceraldehyde-3-phosphate dehydrogenase
1198	Hs.41707	HSPB3	**0.3668**	Hsp27
1203	Hs.83848	TPI1	**0.3395**	Triose phosphate isomerase (TPI)
0523	Hs.65114	KRT18	**0.3335**	Cytokeratin 18
1492	Hs.81915	LAP18	**0.3234**	OP18 (Stathmin)
1493	Hs.81915	LAP18	**0.3154**	OP18 (Stathmin)
1181	Hs.78225	ANXA1	**0.3102**	Annexin variant I
0439	Hs.242463	KRT8	**0.3049**	Cytokeratin 8
0505	Hs.297753	VIM	**0.2939**	Vimentin
0593	Hs.297753	VIM	**0.2809**	Vimentin
1874	Hs.75313	AKR1B1	**0.2790**	Aldose reductase
0935	Hs.75544	YWHAH	**0.2775**	14-3-3 η
2524	Hs.78225	ANXA1	**0.2612**	Annexin I
2324	Hs.65114	KRT18	**0.2601**	Cytokeratin 18
1192	Hs.41707	HSPB3	**0.2558**	Hsp27
0350	Hs.289101	GRP58	**0.2516**	Phospholipase C (GRP58)
0992	Hs.75313	AKR1B1	−0.2460	Aldose reductase
0861	Hs.75313	AKR1B1	0.0761	Aldose reductase
0853	Hs.75313	AKR1B1	−0.0675	Aldose reductase
2503	Hs.76392	ALDH1	−0.0565	Aldehyde dehydrogenase
0381	Hs.76392	ALDH1	−0.0371	Aldehyde dehydrogenase
0371	Hs.76392	ALDH1	−0.0680	Aldehyde dehydrogenase
1179	Hs.78225	ANXA1	0.2052	Annexin variant I
0762	Hs.78225	ANXA1	−0.0739	Annexin I
0760	Hs.78225	ANXA1	−0.0228	Annexin I
2506	Hs.217493	ANXA2	0.2223	Lipocotin (annexin II)
0772	Hs.217493	ANXA2	0.2080	Lipocotin (annexin II)
0723	Hs.217493	ANXA2	0.0701	Lipocotin
1239	Hs.93194	APOA1	0.1133	Apolipoprotein A1 (ApoA1)
1237	Hs.93194	APOA1	−0.0373	Apolipoprotein A1 (ApoA1)
1234	Hs.93194	APOA1	−0.0894	Apolipoprotein A1 (ApoA1)
0428	Hs.25	ATP5B	0.0080	ATP synthase β subunit precursor
0427	Hs.25	ATP5B	0.0122	ATP synthase β subunit precursor
0424	Hs.25	ATP5B	−0.0992	ATP synthase β subunit precursor
0863	Hs.75106	CLU	−0.0483	Apolipoprotein J (ApoJ)
0780	Hs.75106	CLU	−0.0443	Apolipoprotein J (ApoJ)
1527	Hs.119140	EIF5A	−0.0726	eIF-5A
1484	Hs.119140	EIF5A	−0.0376	eIF-5A
1728	Hs.5241	FABP1	−0.1916	L-FABP
1712	Hs.5241	FABP1	−0.0473	L-FABP
0947	Hs.169476	GAPD	0.1745	Glyceraldehyde-3-phosphate dehydrogenase
1232	Hs.75207	GLO1	0.2249	Glyoxalase-I
1229	Hs.75207	GLO1	0.0450	Glyoxalase-1
1595	Hs.158300	HAP1	−0.0137	Huntingtin-associated protein 1 (neuroan 1)
1810	Hs.75990	HP	−0.4672	α-Haptoglobin
1459	Hs.75990	HP	0.0802	α-Haptoglobin
1458	Hs.75990	HP	−0.0305	α-Haptoglobin
0619	Hs.75990	HP	0.0461	B-haptoglobin
0615	Hs.75990	HP	−0.0034	B-haptoglobin
1250	Hs.41707	HSPB3	−0.1024	Hsp27
0549	Hs.79037	HSPD1	0.1074	Hsp60
0338	Hs.79037	HSPD1	0.2265	Hsp60
0333	Hs.79037	HSPD1	0.1383	Hsp60
0331	Hs.79037	HSPD1	0.1603	Hsp60
2381	Hs.65114	KRT18	0.2016	Cytokeratin 18
0535	Hs.65114	KRT18	0.1106	Cytokeratin 18

TABLE II—*continued*

Correlation coefficients of protein and mRNA where multiple isoforms were present on 2D gels

r*, correlation coefficient value > 0.2445; $p < 0.05$. Values in boldface are significant at $p < 0.05$.

Spot	Unigene	Gene name	r*	Protein name
0529	Hs.65114	KRT18	0.1279	Cytokeratin 18
0528	Hs.65114	KRT18	0.0414	Cytokeratin 18
0527	Hs.65114	KRT18	0.0436	Cytokeratin 18
0514	Hs.65114	KRT18	0.0733	Cytokeratin 18
0451	Hs.242463	KRT8	−0.0111	Cytokeratin 8
0446	Hs.242463	KRT8	0.0347	Cytokeratin 8
0444	Hs.242463	KRT8	−0.1311	Cytokeratin 8
0443	Hs.242463	KRT8	0.0942	Cytokeratin 8
1488	Hs.81915	LAP18	0.0495	OP18 (Stathmin)
0321	Hs.75655	P4HB	−0.0546	PDI (proly-4-OH-B)
0320	Hs.75655	P4HB	−0.0041	PDI (proly-4-OH-B)
1063	Hs.75323	PHB	0.0441	Prohibitin
0837	Hs.75323	PHB	0.1402	Prohibitin
0326	Hs.297681	SERPINA1	−0.0227	α-1-Antitripsin
0322	Hs.297681	SERPINA1	−0.0277	α-1-Antitripsin
0241	Hs.297681	SERPINA1	−0.0148	α-1-Antitripsin
1280	Hs.301254	SFTPA1	−0.1488	Pulmonary surfactant-associated protein
1278	Hs.301254	SFTPA1	−0.2040	Pulmonary surfactant-associated protein
0866	Hs.73980	TNNT1	0.1162	Troponin T
0778	Hs.73980	TNNT1	0.0740	Troponin T
1213	Hs.83848	TPI1	0.0024	Triose phosphate isomerase (TPI)
1210	Hs.83848	TPI1	0.0490	Triose phosphate isomerase (TPI)
1207	Hs.83848	TPI1	−0.1615	Triose phosphate isomerase (TPI)
1204	Hs.83848	TPI1	0.0209	Triose phosphate isomerase (TPI)
1202	Hs.83848	TPI1	0.0721	Triose phosphate isomerase (TPI)
1161	Hs.83848	TPI1	0.2265	Triose phosphate isomerase (TPI)
1052	Hs.77899	TPM1	−0.1040	Tropomysin clean-product
1039	Hs.77899	TPM1	−0.2999	Cytoskeletal tropomyosin
1035	Hs.77899	TPM1	−0.3821	Tropomyosin
0783	Hs.77899	TPM1	0.0757	Tropomyosins 1–5
1574	Hs.194366	TTR	−0.0065	Transthyretin
0809	Hs.194366	TTR	0.0399	Transthyretin multimere
2202	Hs.76118	UCHL1	−0.0220	Ubiquitin carboxyl-terminal hydrolase isozyme L1
1246	Hs.76118	UCHL1	−0.1261	Ubiquitin carboxyl-terminal hydrolase isozyme L1
1242	Hs.76118	UCHL1	0.1473	Ubiquitin carboxyl-terminal hydrolase isozyme L1
0606	Hs.297753	VIM	0.0951	Vimentin
0594	Hs.297753	VIM	−0.2664	Vimentin-derived protein (vid4)
0508	Hs.297753	VIM	0.1008	Vimentin-derived protein (vid2)
0419	Hs.297753	VIM	0.0032	Vimentin-derived protein (vid1)
1279	Hs.75544	YWHAH	0.0059	14-3-3 η

such that $\rho_p(i)$ denotes the ith largest correlation coefficient for pth permutation. Hence, the expected correlation coefficient, $\rho_E(i)$, was the average over the 60 permutations, $\rho_E(i) = \Sigma_{p=1}^{60} \rho_p(i)/60$. A scatter plot of observed correlations ($\rho(i)$) *versus* the expected correlations is shown in Fig. 2*D*. For this study, we chose threshold $\Delta = 0.115$ so that correlation would be considered significant if absolute value of difference between $\rho(i)$ and $\rho_E(i)$ was greater than the threshold. Twenty-nine (including one with observed correlation coefficient −0.4672) of 165 pairs of gene and protein expression were called significant in such criteria, and the permuted data generated an average of 5.1 falsely significant pairs of gene and protein expression. This provided an estimated false discovery rate (the percentage of pairs of gene and protein expression identified by chance) for our data set.

RESULTS

Correlation of Individual Proteins and mRNA Expression within Each Tumor—We have examined quantitatively 165 protein spots on 2D gels representing 98 genes and compared protein levels with mRNA levels for a cohort of 85 lung adenocarcinomas and normal lung samples. Of the 165 protein spots, 69 proteins were represented by only one known spot on 2D gels for an individual gene, whereas 96 protein spots showed multiple protein products from 29 different genes. 2D Western blotting verified the proteins identified by mass spectrometry when specific antibodies were available. Spearman correlation coefficients of the proteins and their associated mRNA for each protein spot were generated using all 76 lung adenocarcinomas and nine non-neoplastic lung tissues (see Tables I and II, and see Figs. 1 and 2). The correlation coefficients (r) ranged from −0.467 to 0.442 (Fig. 2*D*). A total of 28 protein spots (21 genes) were found to have a statistically significant correlation between expression of

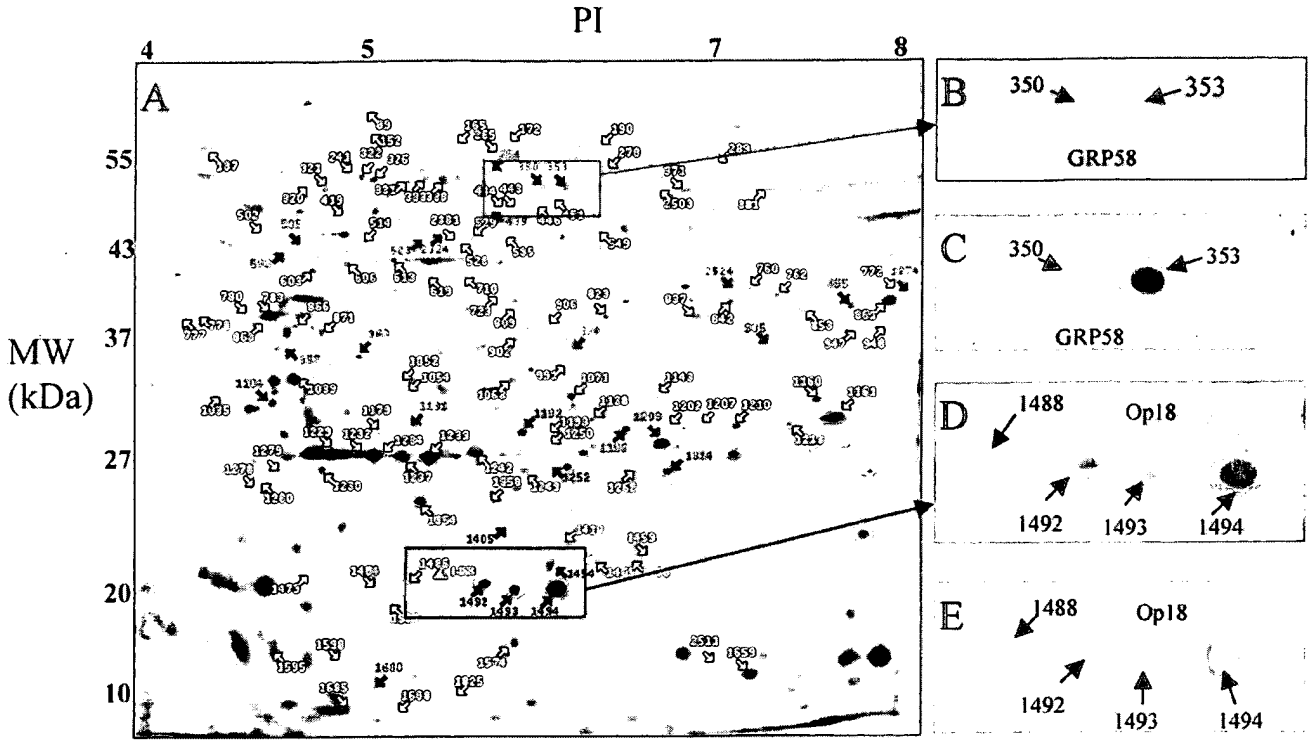


FIG. 1. *A*, digital image of a silver-stained 2D-PAGE separation of a stage I lung adenocarcinoma showing protein spots separated by molecular mass (*MW*) and isoelectric point (*PI*). Twenty-eight protein spots whose expression levels are correlated with mRNA abundance are indicated by the *black arrows*. *B*, the outlined areas of *A* showing protein GRP58. *C*, 2D Western blot of GRP58 from the A549 lung adenocarcinoma cell line. *D*, the outlined areas of *A* showing the protein isoforms of Op18. *E*, 2D Western blot of Op18 from A549 cells.

their protein and mRNA ($r > 0.2445$; $p < 0.05$). This accounts for 17% (28/165) of the 165 protein spots. Among the 69 genes for which only a single protein spot was known (Table I), nine genes (9/69, 13%) were observed to show a statistically significant relationship between protein and mRNA abundance ($r > 0.2445$; $p < 0.05$). The proteins whose expression levels were correlated with their mRNA abundance included those involved in signal transduction, carbohydrate metabolism, apoptosis, protein post-translational modification, structural proteins, and heat shock proteins (Table III).

Individual Isoforms of the Same Protein Have Different Protein/mRNA Correlation Coefficients—Of the 165 protein spots, 96 represent protein products of 29 genes with at least two isoforms. Among these 96 protein spots, 19 (19/96 protein spots, 20%) showed a statistically significant correlation between their protein and mRNA expression ($r > 0.2445$; $p < 0.05$) (Table II) and represented 12 genes (12/29, 41%). Individual isoforms of the same protein demonstrated different protein/mRNA correlation coefficients. For example, 2D-PAGE/Western analysis revealed four isoforms of OP18 differing in regards to isoelectric point but similar in molecular weight. Three of the four isoforms (spots 1492, 1493, and 1494) showed a statistically significant correlation between their protein and mRNA abundance ($r = 0.3234$, 0.3154, and 0.4003, respectively). The forth isoform (spot 1488) showed no correlation between protein and mRNA expression ($r = 0.0495$). Similarly, just one of five quantified isoforms of cytokeratin 8 (spot 439) demonstrated a statistically significant correlation between protein and mRNA abundance ($r = 0.3049$; $p < 0.05$) (Table II).

In addition to differences in the relationship between mRNA levels and protein expression among separate isoforms, some genes with very comparable mRNA levels showed a 24-fold difference in their protein expression. Genes with comparable protein expression levels also showed up to a 28-fold variance in their mRNA levels.

Lack of Correlation for mRNA and Protein Expression when Using Average Tumor Values across All 165 Protein Spots (98 Genes)—The relationship between mRNA and protein expression was also examined by using the average expression values for all samples. To analyze this relationship using this approach, the average value for each protein or mRNA was generated using all 85 lung tissue samples. The range of normalized average protein values ranged from −0.0646 to 0.0979 (raw value 0.0036 to 4.1947), and the range for mRNA was from 0 to 15260.5 for all 165 individual protein spots. The Spearman correlation coefficient for the whole data set (165 protein spots/98 genes) was −0.025 (Fig. 3*A*). Even for the 28 protein spots (Fig. 2*D*) that were found to have a statistically significant correlation between their mRNA and protein, use of

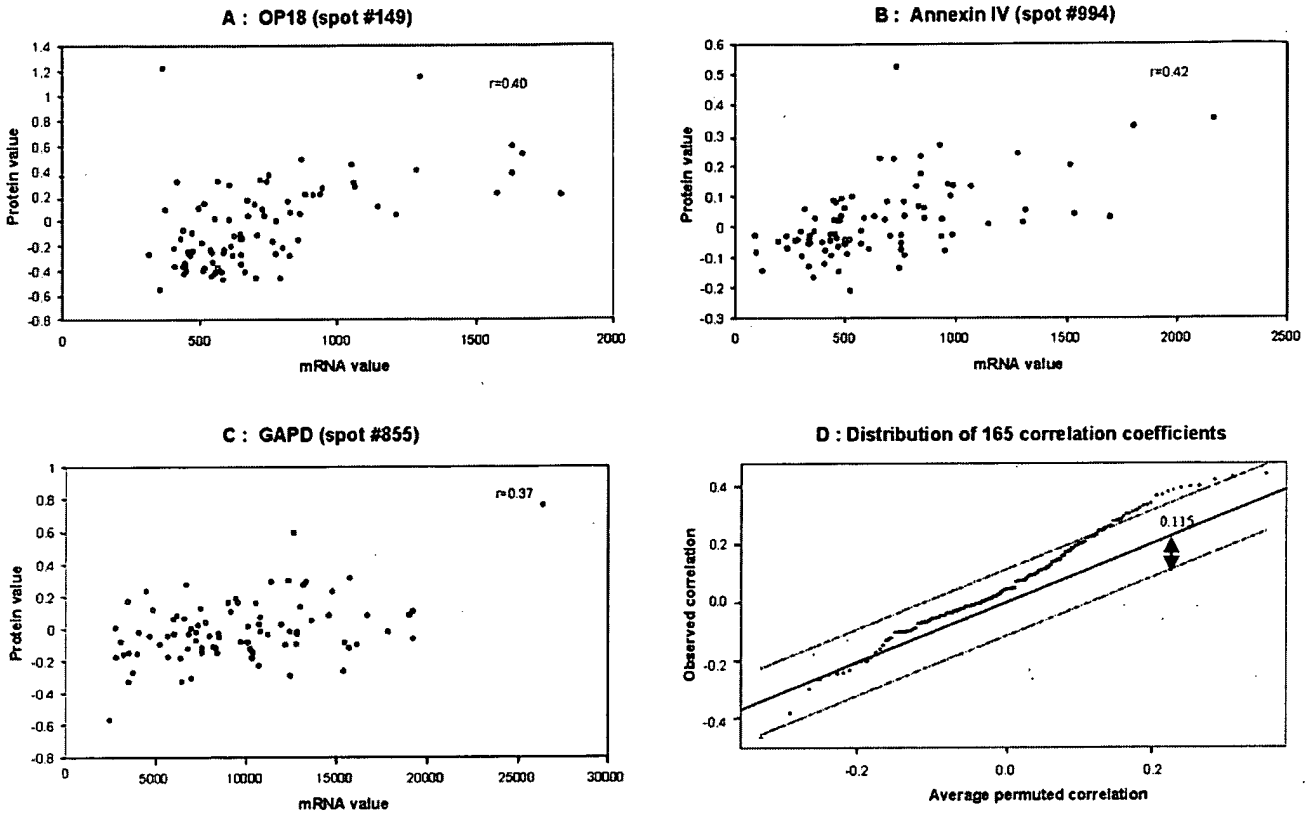


FIG. 2. *A–C*, plots showing the correlation between mRNA and protein for the three selected genes Op18, Annexin IV, and GAPD for all 76 lung adenocarcinomas and nine non-neoplastic lung samples ($p < 0.05$). *D*, distribution of all 165 Spearman correlation coefficients (r) and verification analysis using SAM. A more detailed description of the method is provided under "Experimental Procedures." Approximately 17% of the 165 proteins demonstrate a significant correlation between mRNA and protein levels as demonstrated by the values shown beyond the outer range of threshold $\Delta = 0.115$. Normalized protein values were used, thus negative values for some proteins are observed.

the average value resulted in a correlation coefficient value of −0.035, which was not significant (Fig. 3*B*).

Lack of a Relationship between Protein/mRNA Correlation Coefficients and Average Protein Abundance—To determine whether an absolute protein level might influence the correlation with mRNA, the mean value[s] of each protein (relative abundance) and the Spearman protein/mRNA correlation coefficients among all 85 samples were examined. No relationship between the protein abundance and the correlation coefficients was observed ($r = 0.039$; $p > 0.05$). A detailed analysis of separate subsets of proteins with differing levels of abundance (less than −0.0014, larger than −0.0014, or larger than 0.0077) also showed a lack of correlation between mRNA and protein expression among the 83 (50%), 82 (50%), and 41 (25%) of 165 total protein spots, respectively ($r = 0.016$, 0.08, and 0.172, respectively).

Stage-related Changes in the Protein/mRNA Correlation Coefficients—To determine whether the 21 genes (28 protein spots) showing a significant correlation between the protein and mRNA expression among all samples demonstrate changes in this relationship during tumor progression, the correlations were examined separately for stage I ($n = 57$) and stage III ($n = 19$) lung adenocarcinomas (Table III). The number of non-neoplastic lung samples ($n = 9$) was insufficient for a separate correlation analysis of this group. Many of the protein spots represent one of several known protein isoforms for a given gene. The majority of genes (16/21) did not differ in the protein/mRNA correlation between stage I and stage III tumors indicating a similar regulatory relationship between the mRNA and protein spot. GRP-58, PSMC, SOD1, TPI1, and VIM, however, were found to demonstrate significant differences in the correlation coefficients between stage I and stage III lung adenocarcinomas. For GRP-58, PSMC, and VIM the change in the correlation coefficient was because of a relative increase in protein expression in stage III tumors. For SOD and TPI the change resulted from a relative decrease in expression of this specific protein in stage III tumors.

DISCUSSION

Relatively little is known about the regulatory mechanisms controlling the complex patterns of protein abundance and post-translational modification in tumors. Most reports concerning the regulation of protein translation have focused on

TABLE III
Stage-dependent analysis of protein-mRNA correlation coefficients
r, correlation coefficient. Values in boldface indicate a significant difference between stage I and stage III.

Spot	Gene name	r (Stage I)	r (Stage III)	Function
1874	AKR1B1	0.269	0.106	Carbohydrate metabolism; electron transporter
2524	ANXA1	0.184	0.572	Phospholipase inhibitor; signal transduction
0994	ANXA4	0.660	0.362	Phospholipase inhibitor
0963	ANXA5	0.241	0.390	Phospholipase inhibitor; calcium binding; phospholipid binding
1314	DJ-1	0.363	0.354	Signal transduction
1405	FTL	0.126	0.358	Iron storage protein
0855	GAPD	0.243	0.581	Carbohydrate metabolism (glycolysis regulation)
0350	GRP58	0.327	**−0.087**	Signal transduction; protein disulfide isomerase
0264	HNRPK	0.360	0.243	RNA-binding protein (RNA processing/modification)
1192	HSPB3	0.457	0.633	Heat shock protein
0523	KRT18	0.115	0.371	Structural protein
0439	KRT8	0.323	0.436	Structural protein
1492	LAP18	0.483	0.663	Signal transduction; cell growth and maintenance
1638	LGALS1	0.200	0.528	Apoptosis; cell adhesion; cell size control
1252	PSMC	0.253	**0.060**	Protein degradation
1104	SFN	0.465	0.475	Signal transduction (protein kinase C inhibitor)
1454	SOD1	0.352	**0.079**	Oxidoreductase
1203	TPI1	0.378	**0.009**	Carbohydrate metabolism
0957	TPM1	0.475	0.225	Structural protein (muscle); control of heart
0593	VIM	**−0.054**	0.556	Structural protein
0935	YWHAH	0.283	0.210	Signal transduction

one or several protein products (18). Celis *et al.* (19) found a good correlation between transcript and protein levels among 40 well resolved, abundant proteins using a proteomic and microarray study of bladder cancer. By comparing the mRNA and protein expression levels within the same tumor samples, we found that 17% (28/165) of the protein spots (21/98 genes) show a statistically significant correlation between mRNA and protein. These proteins appear to represent a diverse group of gene products and include those involved in signal transduction, carbohydrate metabolism, protein modification, cell structure, heat shock, and apoptosis. These results suggest that expression of this subset of 165 proteins is likely to be regulated at the transcriptional level in these tissues. The majority of the protein isoforms, however, did not correlate with mRNA levels, and thus their expression is regulated by other mechanisms. We also observed a subset of proteins that demonstrated a negative correlation with the mRNA expression values; for example α-haptoglobin demonstrated a strong negative correlation with its mRNA expression values. This may reflect negative feedback on the mRNA or the protein or the presence of other regulatory influences that are not understood currently.

Post-translational modification or processing will result in individual protein products of the same gene migrating to different locations on 2D-PAGE gels (20). Because the identity of all possible isoforms for each protein examined has not been characterized completely, this may influence the correlation analyses performed in this study. This is partly because of limitations of the 2D-PAGE and mass spectrometry technologies (21, 22). Potential inconsistencies between mRNA and protein correlations that have been reported may also be because of differences, even in the same gene, in the mechanisms of protein translation among different cells or as measured in different laboratories (23).

In this study, we examined 165 protein spots identified in lung adenocarcinomas. Ninety-six protein spots, representing the products of 29 genes, contained at least two protein isoforms. Nineteen of 96 protein spots, representing 12 genes, were shown to have a statistically significant correlation between their protein and mRNA expression, suggesting that the levels of these proteins reflects the transcription of the corresponding genes. Differences in protein/mRNA correlations were found among the individual isoforms of a given protein. For example, of the four OP18 isoforms, three showed a statistically significant correlation between the protein and mRNA expression levels. The lack of relationship for the one isoform, however, indicates that individual protein isoforms of the same gene product can be regulated differentially. This is not unexpected and likely reflects other post-translational mechanisms that can influence isoform abundance in tissues and cancer.

In addition to the analyses of the correlation of mRNA/protein within the same tumor samples, we also tested the global relationship between mRNA and the corresponding protein abundance across all 165 protein spots in the lung samples. A protein and mRNA average value for each gene was generated using all 85 lung tissues samples. We observed a very wide range of normalized average protein and mRNA values. The correlation coefficient generated using this average value data set was −0.025, and even for the 28 protein spots that showed a statistically significant correlation between individual mRNA and proteins, the correlation value was only −0.035. This suggests that it is not possible to predict overall protein expression levels based on average

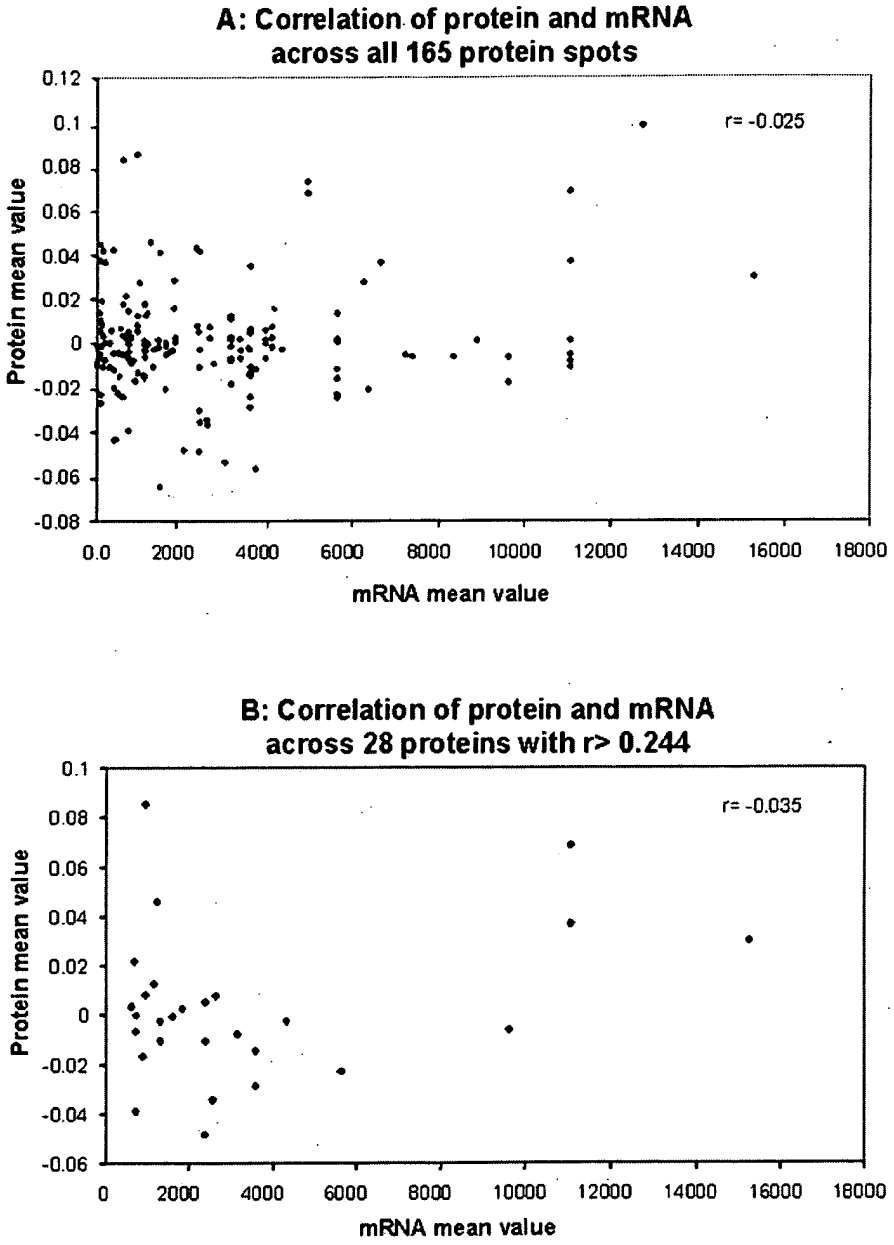


FIG. 3. **The overall correlation of mRNA and protein levels across all 165 protein spots (*A*) and across 28 protein spots that contained individual *r* values larger than 0.244 (*B*) are shown.** Each protein or mRNA mean value was calculated based on all 76 lung adenocarcinomas and nine non-neoplastic lung samples using quantitative 2D-PAGE and Affymetrix oligonucleotide microarrays. The Spearman correlation coefficients for the two data sets (*A* and *B*) were −0.025 and −0.035, respectively, indicating a lack of correlation if mean values for mRNA and protein for all samples is used.

mRNA abundance in lung cancer samples. This conclusion is also supported by previous results from Anderson and Seilhamer (24), who examined 19 genes in human liver cells, and by Gygi *et al.* (25), who examined 106 genes in yeast. Both studies found a lack of correlation between mRNA and protein expression when average or overall levels were used.

A good correlation was reported when the 11 most abundant proteins were examined in yeast (25), suggesting that the level of protein abundance may be a factor that may influence the correlation between mRNA and protein. In the present study, a fairly wide range of mean protein values among 165 protein spots in lung adenocarcinomas was observed, and the correlation coefficients also varied from −0.467 to 0.442. A comparison between the mean value of each protein and the correlation coefficient generated using all 85 tissue samples did not reveal a strong relationship between the overall protein abundance and the correlation coefficients (r = 0.039; $p > 0.05$). Detailed analysis of different subsets of protein abundance also failed to show a correlation between mRNA and protein expression. Thus in contrast to yeast, a relationship between mRNA/protein correlation coefficient and protein abundance in human lung adenocarcinomas was not observed.

The results of this study indicate that the level of protein abundance in lung adenocarcinomas is associated with the corresponding levels of mRNA in 17% (28 proteins) of the total 165 protein spots examined. This was substantially

higher than the amount predicted to result by chance alone (which was 5.1) and suggests that a transcriptional mechanism likely underlies the abundance of these proteins in lung adenocarcinomas. We also demonstrate that the expression of individual isoforms of the same protein may or may not correlate with the mRNA, indicating that separate and likely post-translational mechanisms account for the regulation of isoform abundance. These mechanisms may also account for the differences in the correlation coefficients observed between stage I and stage III tumors, indicating that specific protein isoforms show regulatory changes during tumor progression. Further studies in lung adenocarcinomas will examine the relationship between the expression of individual protein isoforms and specific clinical-pathological features of these tumors, such as the presence of angiolymphatic invasion, and nodal or pleural surface involvement. The potential to identify specific protein isoforms associated with biological behavior in lung adenocarcinomas would be of considerable interest and will add to our understanding of the regulation of gene products by transcriptional, translational, and post-translational mechanisms.

Acknowledgments—We thank Kerby A. Shedden, Rork D. Kuick, Eric Puravs, Robert Hinderer, Melissa C. Krause, and Christopher Wood for assistance in this study.

* This work was supported by NCI, National Institutes of Health Grant U19 CA-85953. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡‡ To whom correspondence should be addressed: General Thoracic Surgery, SRB II, B560, Box 0686, University of Michigan Medical School, Ann Arbor, MI 48109-086; E-mail: dgbeer@umich.edu.

REFERENCES

1. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934
2. Liang, P., Pardee, A. B. (1998) Differential display. A general protocol. *Mol. Biotechnol.* **10,** 261–267
3. Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., Polyak, K. (2001) A sage (serial analysis of gene expression) view of breast tumor progression. *Cancer Res.* **61,** 5697–5702
4. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406,** 536–540
5. Fung, E. T., Wright, G. L., Jr., Dalmasso, E. A. (2000) Proteomic strategies for biomarker identification: progress and challenges. *Curr. Opin. Mol. Ther.* **2,** 643–650
6. Davidson, D., Baldock, R. (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. *Nat. Rev. Genet.* **2,** 409–417
7. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler. J., Lockhart, D. J., Morris, M. S., Fodor, S. P. (1996) Accessing genetic information with high-density DNA arrays. *Science* **274,** 610–614
8. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082
9. Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., Fodor, S. P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5022–5026
10. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 13790–13795
11. Giordano, T. J., Shedden, K. A., Schwartz, D. R., Kuick, R., Taylor, J. M. G., Lee, N., Misek, D. E., Greenson, J. K., Kardia, S. L. R., Beer, D. G., Rennert, G., Cho, K. R., Gruber, S. B., Fearon, E. R., Hanash, S. (2001) Organ-specific molecular classification of lung, colon and ovarian adenocarcinomas using gene expression profiles. *Am. J. Pathol.* **159,** 1231–1238
12. Strahler, J. R., Kuick, R., Hanash, S. M. (1989) in *Protein Structure: A Practical Approach* (Creighton, T., ed) pp. 65–92, IRL Press, Oxford
13. Merril, C. R., Dunau, M. L., Goldman, D. (1981) A rapid sensitive silver stain for polypeptides in polyacrylamide gels. *Anal. Biochem.* **101,** 201–207
14. Hanash, S. M., Strahler, J. R., Chan, Y., Kuick, R., Teichroew, D., Neel, J. V., Hailat, N., Keim, D. R., Gratiot-Deans, J., Ungar, D., Richardson, B. C. (1993) Data base analysis of protein expression patterns during T-cell ontogeny and activation. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 3314–3318
15. Brichory, F. M., Misek, D. E., Yim, A. M., Krause, M. C., Giordano, T. J., Beer, D. G., Hanash, S. M. (2001) An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 9824–9829
16. Lavens-Phillips, S. E., MacGlashan, D. W., Jr. (2000) The tyrosine kinases p53/56lyn and p72syk are differentially expressed at the protein level but not at the messenger RNA level in nonreleasing human basophils. *Am. J. Respir. Cell Mol. Biol.* **23,** 566–571
17. Tusher, V. G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5116–5121
18. Tew, K. D., Monks, A., Barone, L., Rosser, D., Akerman, G., Montali, J. A., Wheatley, J. B., Schmidt, D. E., Jr. (1996) Glutathione-associated enzymes in the human cell lines of the National Cancer Institute Drug Screening Program. *Mol. Pharmacol.* **50,** 149–159
19. Celis, J. E., Kruhoffer, M., Gromova, I., Frederiksen, C., Ostergaard, M., Thykjaer, T., Gromov, P., Yu, J., Palsdottir, H., Magnusson, N., Orntoft, T. F. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* **480,** 2–16
20. Anderson, N. L., Anderson, N. G. (1998) Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19,** 1853–1861
21. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9390–9395
22. Fey, S. J., Larsen, P. M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.* **5,** 26–33
23. McBride, S., Walsh, D., Meleady, P., Daly, N., Clynes, M. (1999) Bromodeoxyuridine induces keratin protein synthesis at a posttranscriptional level in human lung tumor cell lines. *Differentiation* **64,** 185–193
24. Anderson, L., Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18,** 533–537
25. Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730

(or perhaps was exacerbated by) a UK government seen to be welcoming of GM foods and crops. Another negative was that it was major transnational corporations—another questionable community in the eyes of much of the public here—that were seeking to push their new products onto the public without previous debate and without there being any perceptible benefit. And finally, the potentially negative impact of GM crops on organic farmers—who are seen by some as crucially important for the sustainable future of food production—and the relatively small scale of agricultural production in the United Kingdom (and Europe) have also been important issues.

The question to be answered, therefore, is not how to force the EU to accept GM foods and crops against its own public opinion, but how to change public opinion in the EU. The UK government is currently conducting several exercises that it hopes will provide the facts to support a relaxation of the moratorium on growing GM crops. These include a major review of the costs and benefits of GM crops (just finished), a scientific review of the issues (also now finished), a series of crop trials (results in September) and a public debate on GM crops, 'GM nation' (just finished).

Whether these will change attitudes is moot: the costs-and-benefits review has concluded that the economic value of the few currently available GM crops that could be grown in the UK is likely to be limited because of negative consumer attitudes to GM foods.

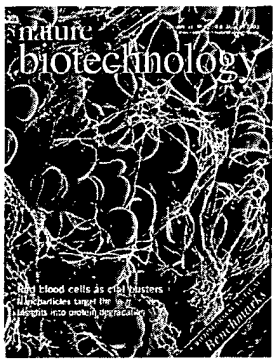
M H Barnes

3 The Spinney,
Watford WD17 4QF, UK
e-mail: MBarnes466@aol.com

To the editor:

Several articles in the July and August issues of *Nature Biotechnology* (21, 735–738, 2003; 21, 852–854, 2003) discuss whether the US strategy of forcing the European Union (EU; Brussels, Belgium) to accept GM foods by referring to World Trade Organisation (WTO; Geneva, Switzerland) rules will bear fruit. We do not believe so—rather the opposite.

A central claim in the arguments of both President Bush and US commerce representative Robert B. Zoellick is that the risk of GM foods is negligible. The veracity of that statement, however, depends on what is defined as risk. A common understanding is that risk relates to the environment and human health. On the other hand, recent studies have repeatedly shown that public hesitance also includes a number of ethical issues (*e.g.*, market dominance of a few large companies and GM crops threatening natural or divine orders, refs 1,2). Our worry is that the US government is neglecting widespread concerns of the European public that include more than environmental risk and human health.



Research carried out by our group in Denmark[1] indicates that, although many people are confident that the public authorities are able to manage the risks here and now, people are less confident about their ability to handle long-term effects because of the scientific uncertainty. Attempts to conceal these or other limits to scientific knowledge do not prevent controversies from arising; rather, the opposite happens because trust in business, scientific experts and public authorities is undermined (witness the handling of the BSE controversy in the United Kingdom).

In the long run, a policy of openness about the different dimensions of uncertainty would be more likely to increase trust in scientific risk assessment. Of course, this will not guarantee public acceptance of GM food, but experience in Europe shows that transparency and dialog are prerequisites for decreasing concerns about new technology.

The argument that the EU's resistance to GM food has had negative consequences for developing countries, denying them access to a technology that could alleviate food provision, is regarded sympathetically by many among the European public. Indeed, here most people abandon the simple dichotomy between 'unacceptable' GM food and the much more acceptable medical uses. This is because GM foods are seen as a means to help people in distress. Many counter such humanitarian uses, however, by the observation that, in general, GM crops are developed not to benefit people in the developing world, but to make money. Needless to say, according to those who point this out, making money is not in itself an acceptable objective. Thus, the fear is that the benefits will never accrue to those who are at present suffering.

Kristian Borch, Jesper Lassen & Rikke B Jørgensen

Centre for Bioethics and Risk Assessment, Systems Analysis Department, PO Box 49, DK-4000 Roskilde, Denmark
e-mail: kristian.borch@risoe.dk

1. Lassen, J. *et al. Bioprocess Biosyst. Eng.* **24**, 263–271 (2002).
2. Wagner, W. *et al.* in *Biotechnology 1996–2000. The Years of Controversy* (Gaskel, G. & Baver, M.W., eds.) 80–95 (Science Museum, London, 2001)

Mining the literature and large datasets

To the editor:

In the accelerating quest for disease biomarkers, the use of high-throughput technologies, such as DNA microarrays and proteomics experiments, has produced vast datasets identifying thousands of genes whose expression patterns differ in diseased versus normal samples. Although many of these differences may reach statistical significance, they are not always biologically meaningful. For example, reports of mRNA or protein changes of as little as two-fold are not uncommon, and although some changes of this magnitude turn out to be important, most are attributable to disease-independent differences between the samples. Evidence gleaned from other studies linking genes to the disease is helpful, but with such large datasets, a manual literature review is often not practical. Thus, the power of these emerging technologies—the ability to quickly generate large sets of data—has challenged current means of evaluating and validating these data. One study from 1999, for example, reveals that a researcher would have to scan 130 different journals and read 27 papers per day to follow a single disease, such as breast cancer[1].

To address this need, my group at

Harvard recently developed a freely accessible automated literature-mining tool, termed MedGene, that comprehensively summarizes the relationships among over 50,000 named human genes (and their synonyms) and over 4,000 human diseases from over 12 million records in Medline (http://hipseq.med.harvard.edu/MedGene). Several key features of this resource are worth noting. First, MedGene is not limited to any specific relationship type, but rather encompasses all reported gene-disease links, including the genetic, biochemical, pharmacological, epidemiological and physiological. Second, the database assigns a mathematical score summarizing the strength of the association between the disease and the gene, which allows semiquantitative analysis and organizes the genes in rank order. Finally, the relationships are identified automatically by advanced text searching and filtering algorithms that result in low rates of false-positive and false-negative linkages[2]. In one query, MedGene identified nearly 2,400 breast cancer–related genes, whereas the same search in four commonly used databases yielded a combined total of 286 genes, 260 of which were included in the MedGene list[1–5].

A summary of all gene-disease relationships offers the unique opportunity to both evaluate and validate the outcome of high-throughput experiments. For example, we used MedGene to analyze a DNA microarray experiment in which over 2,000 genes demonstrated statistically significant differences in expression between normal breast tissue and breast cancer. It was able to identify the subset of these genes previously described as breast cancer–related genes in the literature. To determine whether gene expression level correlated with the strength of the association between gene and breast cancer, we plotted gene expression levels against the breast cancer literature relationship scores assigned by MedGene. Interestingly, there is no correlation when considering expression differences as high as fivefold; however, a significant correlation is observed ($r = 0.41$; $P = 0.05$) among genes showing a difference of tenfold or more. Thus, for this experiment, expression level differences as high as fivefold cannot be attributed to the disease without corroborating evidence. It will be interesting to learn if similar results hold for other diseases and other experiments.

As the search for disease biomarkers and drug targets comes to rely increasingly upon genomic-scale technologies, demand will grow for automated resources, such as MedGene, that help process the resulting data volume.

Joshua LaBaer

Institute of Proteomics,
Harvard Medical School,
250 Longwood Ave., BCMP,
Boston, Massachusetts 02115, USA
e-mail: josh@hms.harvard.edu

1. Baasiri, R.A., Glasser, S.R., Steffen, D.L. & Wheeler, D.A. *Oncogene* **18**, 7958–7965 (1999).
2. Hu, Y. *et al. J. Proteome Res.* **2**, 405–412 (2003).
3. Steffen, D.L., Levine, A.E., Yarus, S., Baasiri, R.A. & Wheeler, D.A. *Bioinformatics* **16**, 639–649 (2000).
4. Bairoch, A. & Apweiler, R. *Nucleic Acids Res.* **28**, 45–48 (2000).
5. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. & Lancet, D. *Trends Genet.* **13**, 163 (1997).

MOLECULAR AND CELLULAR BIOLOGY, Mar. 1999, p. 1720–1730 Vol. 19, No. 3
0270-7306/99/$04.00+0

Correlation between Protein and mRNA Abundance in Yeast

STEVEN P. GYGI, YVAN ROCHON, B. ROBERT FRANZA, AND RUEDI AEBERSOLD*

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730

Received 5 October 1998/Returned for modification 11 November 1998/Accepted 2 December 1998

We have determined the relationship between mRNA and protein expression levels for selected genes expressed in the yeast *Saccharomyces cerevisiae* growing at mid-log phase. The proteins contained in total yeast cell lysate were separated by high-resolution two-dimensional (2D) gel electrophoresis. Over 150 protein spots were excised and identified by capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS). Protein spots were quantified by metabolic labeling and scintillation counting. Corresponding mRNA levels were calculated from serial analysis of gene expression (SAGE) frequency tables (V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler, Cell 88:243–251, 1997). We found that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Indeed, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold. Another interesting observation is that codon bias is not a predictor of either protein or mRNA levels. Our results clearly delineate the technical boundaries of current approaches for quantitative analysis of protein expression and reveal that simple deduction from mRNA transcript analysis is insufficient.

The description of the state of a biological system by the quantitative measurement of the system constituents is an essential but largely unexplored area of biology. With recent technical advances including the development of differential display-PCR (21), of cDNA microarray and DNA chip technology (20, 27), and of serial analysis of gene expression (SAGE) (34, 35), it is now feasible to establish global and quantitative mRNA expression profiles of cells and tissues in species for which the sequence of all the genes is known. However, there is emerging evidence which suggests that mRNA expression patterns are necessary but are by themselves insufficient for the quantitative description of biological systems. This evidence includes discoveries of posttranscriptional mechanisms controlling the protein translation rate (15), the half-lives of specific proteins or mRNAs (33), and the intracellular location and molecular association of the protein products of expressed genes (32).

Proteome analysis, defined as the analysis of the protein complement expressed by a genome (26), has been suggested as an approach to the quantitative description of the state of a biological system by the quantitative analysis of protein expression profiles (36). Proteome analysis is conceptually attractive because of its potential to determine properties of biological systems that are not apparent by DNA or mRNA sequence analysis alone. Such properties include the quantity of protein expression, the subcellular location, the state of modification, and the association with ligands, as well as the rate of change with time of such properties. In contrast to the genomes of a number of microorganisms (for a review, see reference 11) and the transcriptome of *Saccharomyces cerevisiae* (35), which have been entirely determined, no proteome map has been completed to date.

The most common implementation of proteome analysis is the combination of two-dimensional gel electrophoresis (2DE) (isoelectric focusing-sodium dodecyl sulfate [SDS]-polyacrylamide gel electrophoresis) for the separation and quantitation of proteins with analytical methods for their identification. 2DE permits the separation, visualization, and quantitation of thousands of proteins reproducibly on a single gel (18, 24). By itself, 2DE is strictly a descriptive technique. The combination of 2DE with protein analytical techniques has added the possibility of establishing the identities of separated proteins (1, 2) and thus, in combination with quantitative mRNA analysis, of correlating quantitative protein and mRNA expression measurements of selected genes.

The recent introduction of mass spectrometric protein analysis techniques has dramatically enhanced the throughput and sensitivity of protein identification to a level which now permits the large-scale analysis of proteins separated by 2DE. The techniques have reached a level of sensitivity that permits the identification of essentially any protein that is detectable in the gels by conventional protein staining (9, 29). Current protein analytical technology is based on the mass spectrometric generation of peptide fragment patterns that are idiotypic for the sequence of a protein. Protein identity is established by correlating such fragment patterns with sequence databases (10, 22, 37). Sophisticated computer software (8) has automated the entire process such that proteins are routinely identified with no human interpretation of peptide fragment patterns.

In this study, we have analyzed the mRNA and protein levels of a group of genes expressed in exponentially growing cells of the yeast *S. cerevisiae*. Protein expression levels were quantified by metabolic labeling of the yeast proteins to a steady state, followed by 2DE and liquid scintillation counting of the selected, separated protein species. Separated proteins were identified by in-gel tryptic digestion of spots with subsequent analysis by microspray liquid chromatography-tandem mass spectrometry (LC-MS/MS) and sequence database searching. The corresponding mRNA transcript levels were calculated from SAGE frequency tables (35).

This study, for the first time, explores a quantitative comparison of mRNA transcript and protein expression levels for a relatively large number of genes expressed in the same metabolic state. The resultant correlation is insufficient for predic-

* Corresponding author. Mailing address: Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730. Phone: (206) 221-4196. Fax: (206) 685-7301. E-mail: ruedi@u.washington.edu.

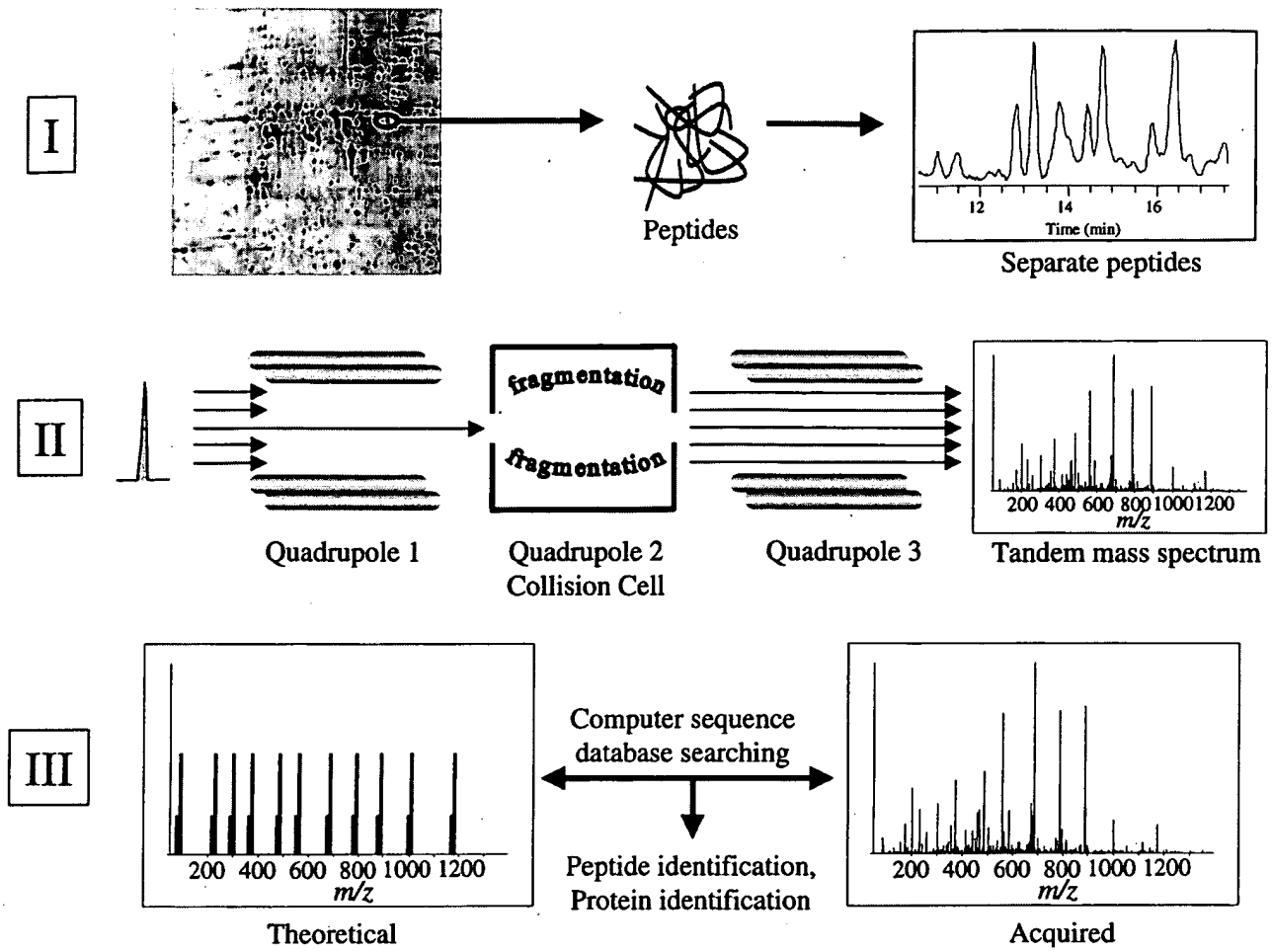


FIG. 1. Schematic illustration of proteome analysis by 2DE and mass spectrometry. In part I, proteins are separated by 2DE, stained spots are excised and subjected to in-gel digestion with trypsin, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. In part II, a peptide is shown eluting from the column in part I. The peptide is ionized by electrospray ionization and enters the mass spectrometer. The mass of the ionized peptide is detected, and the first quadrupole mass filter allows only the specific mass-to-charge ratio of the selected peptide ion to pass into the collision cell. In the collision cell, the energized, ionized peptides collide with neutral argon gas molecules. Fragmentation of the peptide is essentially random but occurs mainly at the peptide bonds, resulting in smaller peptides of differing lengths (masses). These peptide fragments are detected as a tandem mass (MS/MS) spectrum in the third quadrupole mass filter where two ion series are recorded simultaneously, one each from sequencing inward from the N and C termini of the peptide, respectively. In part III, the MS/MS spectrum from the selected, ionized peptide is compared to predicted tandem mass spectra computer generated from a sequence database. Provided that the peptide sequence exists in the database, the peptide and, by association, the protein from which the peptide was derived can be identified. Unambiguous protein identification is attained in a single analysis because multiple peptides are identified as being derived from the same protein.

tion of protein levels from mRNA transcript levels. We have also compared the relative amounts of protein and mRNA with the respective codon bias values for the corresponding genes. This comparison indicates that codon bias by itself is insufficient to accurately predict either the mRNA or the protein expression levels of a gene. In addition, the results demonstrate that only highly expressed proteins are detectable by 2DE separation of total cell lysates and that therefore the construction of complete proteome maps with current technology will be very challenging, irrespective of the type of organism.

MATERIALS AND METHODS

Yeast strain and growth conditions. The source of protein and message transcripts for all experiments was YPH499 (*MATa ura3-52 lys2-801 ade2-101 leu2-Δ1 his3-Δ200 trp1-Δ63*) (30). Logarithmically growing cells were obtained by growing yeast cells to early log phase (3×10^6 cells/ml) in YPD rich medium (YPD supplemented with 6 mM uracil, 4.8 mM adenine, and 24 mM tryptophan) at 30°C (35). Metabolic labeling of protein was accomplished in YPD medium exactly as described elsewhere (4) with the exception that 1 ml of cells was labeled with 3 mCi to offset methionine present in YPD medium. Protein was harvested as described by Garrels and coworkers (12). Harvested protein was lyophilized, resuspended in isoelectric focusing gel rehydration solution, and stored at −80°C.

2DE. Soluble proteins were run in the first dimension by using a commercial flatbed electrophoresis system (Multiphor II; Pharmacia Biotech). Immobilized polyacrylamide gel (IPG) dry strips with nonlinear pH 3.0 to 10.0 gradients (Amersham-Pharmacia Biotech) were used for the first-dimension separation. Forty micrograms of protein from whole-cell lysates was mixed with IPG strip rehydration buffer (8 M urea, 2% Nonidet P-40, 10 mM dithiothreitol), and 250 to 380 μl of solution was added to individual lanes of an IPG strip rehydration tray (Amersham-Pharmacia Biotech). The strips were allowed to rehydrate at room temperature for 1 h. The samples were run at 300 V–10 mA–5 W for 2 h, then ramped to 3,500 V–10 mA–5 W over a period of 3 h, and then kept at 3,500 V–10 mA–5 W for 15 to 19 h. At the end of the first-dimension run (60 to 70 kV · h), the IPG strips were reequilibrated for 8 min in 2% (wt/vol) dithiothreitol in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8) and for 4 min in 2.5% iodoacetamide in 2% (wt/vol) SDS–6 M urea–30% (wt/vol) glycerol–0.05 M Tris HCl (pH 6.8). Following reequilibration, the strips were transferred and apposed to 10% polyacrylamide second-dimension gels. Polyacrylamide gels were poured in a casting stand with 10% acrylamide–2.67% piperazine diacrylamide–0.375 M Tris base-HCl (pH 8.8)–0.1% (wt/vol) SDS–0.05%

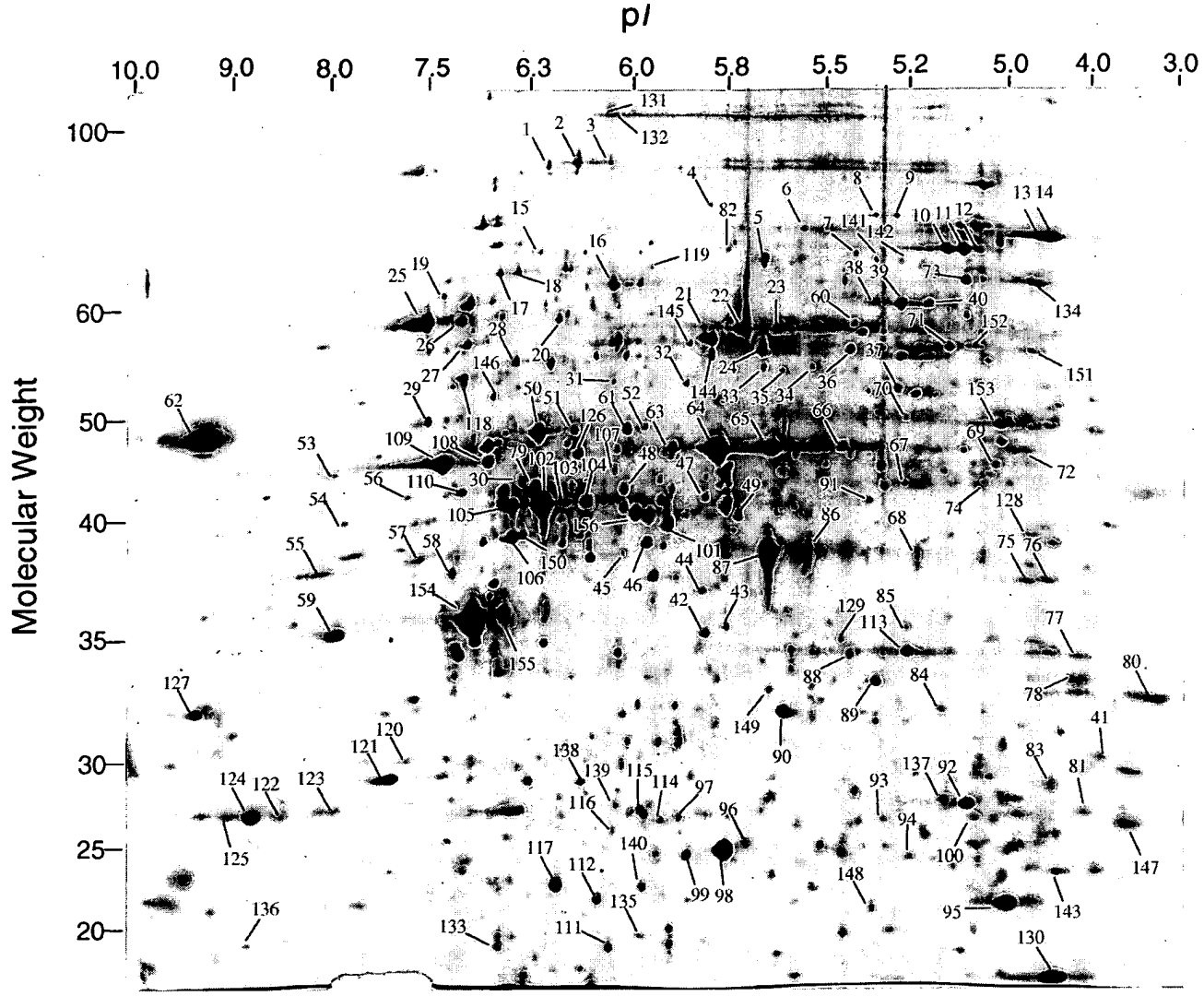


FIG. 2. 2D silver-stained gel of the proteins in yeast total cell lysate. Proteins were separated in the first dimension (horizontal) by isoelectric focusing and then in the second dimension (vertical) by molecular weight sieving. Protein spots (156) were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. Spots were excised, and the corresponding protein was identified by mass spectrometry and database searching. The spots are labeled on the gel and correspond to the data presented in Table 1. Molecular weights are given in thousands.

(wt/vol) ammonium persulfate–0.05% TEMED (*N,N,N′,N′*-tetramethylethylenediamine) in Milli-Q water. The apparatus used to run second-dimension gels was a noncommercial apparatus from Oxford Glycosciences, Inc. Once the IPG strips were apposed to the second-dimension gels, they were immediately run at 50 mA (constant)–500 V–85 W for 20 min, followed by 200 mA (constant)–500 V–85 W until the buffer front line was 10 to 15 mm from the bottom of the gel. Gels were removed and silver stained according to the procedure of Shevchenko et al. (29).

Protein identification. Gels were exposed to X-ray film overnight, and then the silver staining and film were used to excise 156 spots of varying intensities, molecular weights, and isoelectric focusing points. In order to increase the detection limit by mass spectrometry, spots were cut out and pooled from up to four identical cold, silver-stained gels. In-gel tryptic digests of pooled spots were performed as described previously (29). Tryptic peptides were analyzed by microcapillary LC-MS with automated switching to MS/MS mode for peptide fragmentation. Spectra were searched against the composite OWL protein sequence database (version 30.2; 250,514 protein sequences) (24a) by using the computer program Sequest (8), which matches theoretical and acquired tandem mass spectra. A protein match was determined by comparing the number of peptides identified and their respective cross-correlation scores. All protein identifications were verified by comparison with theoretical molecular weights and isoelectric points.

mRNA quantitation. Velculescu and coworkers have previously generated frequency tables for yeast mRNA transcripts from the same strain grown under the same stated conditions as described herein (35). The SAGE technology is based on two main principles. First, a short sequence tag (15 bp) that contains sufficient information uniquely to identify a transcript is generated. A single tag is usually generated from each mRNA transcript in the cell which corresponds to 15 bp at the 3′-most cutting site for *Nla*III. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously. Over 20,000 transcripts were sequenced from yeast strain YPH499 growing at mid-log phase on glucose. Assuming the previously derived estimate of 15,000 mRNA molecules per cell (16), this would represent a 1.3-fold coverage even for mRNA molecules present at a single copy per cell and would provide a 72% probability of detecting such transcripts. Computer software which took for input the gene detected, examined the nucleotide sequence, and performed the calculation as described by Velculescu and coworkers (35) was written. In practice, we found that for 21 of 128 (16%) genes examined viable mRNA levels from SAGE data could not be calculated. This was because (i) no CATG site was found in the open reading frame (ORF), (ii) a CATG site was found but the corresponding 10-bp putative SAGE tag was not found in the frequency tables, or (iii) identical putative SAGE tags were present for multiple genes (e.g., TDH2_YEAST and TDH3_YEAST).

TABLE 1. Expressed genes identified from 2D gel in Fig. 2

Mol wt	pI	Spot no.	YPD gene name[a]	Protein abundance (10^3 copies/cell)	mRNA abundance (copies/cell)	Codon bias
17,259	6.75	133	CPR1	15.2	61.7	0.769
18,702	4.80	83	EGD2	20.1	5.2	0.724
18,726	4.44	147	YKL056C	61.2	88.4	0.831
18,978	5.95	135	YER067W	3.7	6.7	0.118
19,108	5.04	130	YLR109W	94.4	9.7	0.680
19,681	9.08	136	ATP7	11.0	NA[b,c]	0.246
20,505	6.07	111	GUK1	16.5	3.7	0.422
21,444	5.25	148	SAR1	5.4	10.4	0.455
21,583	4.98	95	TSA1	110.6	40.1	0.845
22,602	4.30	80	EFB1	66.1	23.8	0.875
23,079	6.29	112	SOD2	12.6	2.2	0.351
23,743	5.44	137	HSP26	NA[d]	0.7	0.434
24,033	5.97	96	ADK1	17.4	16.4	0.656
24,058	4.43	143	YKL117W	29.2	10.4	0.339
24,353	6.30	140	TFS1	8.1	0.7	0.146
24,662	5.85	99	URA5	25.4	6.0	0.359
24,808	6.33	97	GSP1	26.3	5.2	0.735
24,908	8.73	122	RPS5	18.6	NA[c]	0.899
25,081	4.65	81	MRP8	9.3	NA[c]	0.241
25,960	6.06	116	RPE1	5.8	0.7	0.372
26,378	9.55	127	RPS3	96.8	NA[c]	0.863
26,467	5.18	100	VMA4	10.5	3.7	0.427
26,661	5.84	98	TPI1	NA[d]	NA[c]	0.900
27,156	5.56	93	PRE8	6.9	0.7	0.129
27,334	6.13	115	YHR049W	18.4	2.2	0.520
27,472	5.33	92	YNL010W	31.6	3.7	0.421
27,480	8.95	123	GPM1	10.0	169.4	0.902
27,480	8.95	124	GPM1	231.4	169.4	0.902
27,480	8.95	125	GPM1	7.5	169.4	0.902
27,809	5.97	139	HOR2	5.7	0.7	0.381
27,874	4.46	78	YST1	13.6	52.8	0.805
28,595	4.51	41	PUP2	4.4	0.7	0.147
29,156	6.59	114	YMR226C	14.5	2.2	0.283
29,244	8.40	120	DPM1	5.0	11.2	0.362
29,443	5.91	48	PRE4	3.4	3.7	0.162
30,012	6.39	138	PRB1	21.2	1.5	0.449
30,073	4.63	77	BMH1	14.7	28.2	0.454
30,296	7.94	121	OMP2	67.4	41.6	0.499
30,435	6.34	89	GPP1	70.2	11.2	0.703
31,332	5.57	88	ILV6	13.9	3.0	0.402
32,159	5.46	113	IPP1	63.1	3.7	0.752
32,263	6.00	149	HIS1	22.4	4.5	0.232
33,311	5.35	84	SPE3	15.1	6.7	0.468
34,465	5.60	129	ADE1	8.7	5.2	0.305
34,762	5.32	85	SEC14	10.9	6.0	0.373
34,797	5.85	42	URA1	49.5	8.9	0.237
34,799	6.04	90	BEL1	103.2	81.0	0.875
35,556	5.97	43	YDL124W	6.4	4.5	0.206
35,619	8.41	59	TDH1	69.8	32.7[c]	0.940
35,650	5.49	68	CAR1	5.2	3.0	0.339
35,712	6.72	117	TDH2	49.6	473.0[c]	0.982
35,712	6.72	154	TDH2	863.5	473.0[c]	0.982
35,712	6.72	155	TDH2	79.4	473.0[c]	0.982
36,272	4.85	128	APA1	8.7	0.7	0.425
36,358	5.05	75	YJR105W	17.6	17.1	0.522
36,358	5.05	76	YJR105W	27.5	17.1	0.522
36,596	6.37	79	ADH2	58.9	260.0[c]	0.711
36,714	6.30	102	ADH1	746.1	260.0	0.913
36,714	6.30	103	ADH1	17.6	260.0	0.913
36,714	6.30	104	ADH1	61.4	260.0	0.913
36,714	6.30	105	ADH1	52.7	260.0	0.913
37,033	6.23	44	TAL1	44.8	3.7	0.701
37,796	7.36	57	IDH2	29.4	6.7	0.330
37,886	6.49	106	ILV5	76.0	4.5	0.892
38,700	7.83	55	BAT1	30.9	11.2	0.469
38,702	6.24	46	QCR2	NA[d]	2.2	0.326
39,477	5.58	86	FBA1	17.8	183.6	0.935
39,477	5.58	87	FBA1	427.2	183.6	0.935
39,540	6.50	150	HOM2	60.3	4.5	0.592
39,561	6.12	156	PSA1	96.4	27.5	0.718
41,158	6.01	49	YNL134C	14.9	1.5	0.316
41,623	7.18	58	BAT2	19.0	8.9	0.250
41,728	7.29	110	ERG10	24.1	4.5	0.543
41,900	5.42	74	TOM40	22.3	2.2	0.375
42,402	6.29	45	CYS3	6.7	8.9	0.621
42,883	5.63	67	DYS1	15.8	5.2	0.526
43,409	6.31	107	SER1	10.5	1.5	0.292
43,421	5.59	91	ERG6	2.2	14.1	0.408
44,174	7.32	56	YBR025C	13.1	6.0	0.684
44,682	4.99	72	TIF1	2.9	39.4	0.834
44,707	7.77	108	PGK1	23.7	165.7	0.897
44,707	7.77	109	PGK1	315.2	165.7	0.897
46,080	6.72	30	CAR2	15.4	NA[c]	0.495
46,383	8.52	53	IDP1	7.7	0.7	0.436
46,553	5.98	47	IDP2	32.4	NA[c]	0.197
46,679	6.39	50	ENO1	35.4	0.7	0.930
46,679	6.39	51	ENO1	6.6	0.7	0.930
46,679	6.39	52	ENO1	2.2	0.7	0.930
46,773	5.82	63	ENO2	15.5	289.1	0.960
46,773	5.82	64	ENO2	635.5	289.1	0.960
46,773	5.82	65	ENO2	93.0	289.1	0.960
46,773	5.82	66	ENO2	31.0	289.1	0.960
47,402	6.09	126	COR1	2.5	0.7	0.422
47,666	8.98	54	AAT2	11.7	6.0	0.338
48,364	5.25	73	WTM1	74.5	13.4	0.365
48,530	6.20	61	MET17	38.1	29.0	0.576
48,904	5.18	69	LYS9	16.2	3.7	0.463
48,987	4.90	153	SUP45	29.6	11.9	0.377
49,727	5.47	70	PRO2	13.6	5.2	0.297
49,912	9.27	62	TEF2	558.5	282.0	0.932
50,444	5.67	35	YDR190C	4.8	2.2	0.228
50,837	6.11	32	YEL047C	3.8	1.5	0.387
50,891	4.59	151	TUB2	11.2	7.4	0.404
51,547	6.80	27	LPD1	18.9	2.2	0.351
52,216	7.25	29	SHM2	19.7	7.4	0.722
52,859	5.54	37	YFR044C	30.2	6.7	0.442
53,798	5.19	71	HXK2	26.5	7.4	0.756
53,803	6.05	145	GYP6	4.4	0.7	0.147
54,403	5.29	39	ALD6	37.7	2.2	0.664
54,403	5.29	40	ALD6	6.6	2.2	0.664
54,502	6.20	31	ADE13	6.3	1.5	0.417
54,543	7.75	25	PYK1	225.3	101.8	0.965
54,543	7.75	26	PYK1	39.8	101.8	0.965
55,221	6.66	146	YEL071W	16.3	3.0	0.244
55,295	4.35	134	PDI1	66.2	14.1	0.589
55,364	5.98	24	GLK1	22.6	6.0	0.237
55,481	7.97	118	ATP1	21.6	2.2	0.637
55,886	6.47	28	CYS4	22.2	NA[c]	0.444
56,167	5.83	33	ARO8	14.3	3.0	0.324
56,167	5.83	34	ARO8	9.1	3.0	0.324
56,584	6.36	20	CYB2	18.9	NA[c]	0.259
57,366	5.53	60	FRS2	2.3	0.7	0.451
57,383	5.98	144	ZWF1	5.6	0.7	0.215
57,464	5.49	36	THR4	21.4	3.7	0.508
57,512	5.50	7	SRV2	6.5	NA[c]	0.260
57,727	4.92	152	VMA2	33.7	8.9	0.546
58,573	6.47	17	ACH1	4.4	1.5	0.327
58,573	6.47	18	ACH1	5.4	1.5	0.327
61,353	5.87	21	PDC1	6.5	200.7	0.962
61,353	5.87	22	PDC1	303.2	200.7	0.962
61,353	5.87	23	PDC1	16.3	200.7	0.962
61,649	5.54	38	CCT8	2.2	1.5	0.271

Continued

Continued on following page

TABLE 1—*Continued*

Mol wt	pI	Spot no.	YPD gene name[a]	Protein abundance (10^3 copies/cell)	mRNA abundance (copies/cell)	Codon bias
61,902	6.21	101	PDC5	4.3	NA[c]	0.828
62,266	6.19	16	ICL1	20.1	NA[c]	0.327
62,862	8.02	19	ILV3	5.3	4.5	0.548
63,082	6.40	119	PGM2	2.2	3.0	0.402
64,335	5.77	5	PAB1	30.4	1.5	0.616
66,120	5.42	8	STI1	6.7	0.7	0.313
66,120	5.42	9	STI1	6.4	0.7	0.313
66,450	5.29	141	SSB2	7.0	NA[c]	0.880
66,450	5.29	142	SSB2	2.3	NA[c]	0.880
66,456	5.23	10	SSB1	64.5	79.5	0.907
66,456	5.23	11	SSB1	59.0	79.5	0.907
66,456	5.23	12	SSB1	13.7	79.5	0.907
68,397	5.82	82	LEU4	3.1	3.0	0.407
69,313	4.90	13	SSA2	24.3	18.6	0.892
69,313	4.90	14	SSA2	77.1	18.6	0.892
74,378	8.46	15	YKL029C	2.8	3.7	0.353
75,396	5.82	6	GRS1	5.5	7.4	0.500
85,720	6.25	1	MET6	2.0	NA[c]	0.772
85,720	6.25	2	MET6	10.9	NA[c]	0.772
85,720	6.25	3	MET6	1.4	NA[c]	0.772
93,276	6.11	131	EFT1	17.9	41.6	0.890
93,276	6.11	132	EFT1	5.7	41.6	0.890
102,064[e]	6.61[e]	94	ADE3	4.8	5.2	0.423
107,482[e]	5.33[e]	4	MCM3	2.7	NA[c]	0.240

[a] YPD gene names are available from the YPD website (39).
[b] NA, calculation could not be performed or was not available.
[c] mRNA data inconclusive or NA.
[d] No methionines in predicted ORF; therefore, protein concentration was not determined.
[e] Measured molecular weight or pI did not match theoretical molecular weight or pI.

Protein quantitation. [^{35}S]methionine-labeled gels were exposed to X-ray film overnight, and then the silver stain and film were used to excise 156 spots of varying intensities, molecular weights, and pIs. The excised spots were placed in 0.6-ml microcentrifuge tubes, and scintillation cocktail (100 μl) was added. The samples were vortexed and counted. In addition, two parallel gels were electroblotted to polyvinylidene difluoride membranes. The membranes were exposed to X-ray film, and four intense single spots were excised from each membrane and subjected to amino acid analysis. For these four spots, a mean of 209 ± 4 cpm/pmol of protein/methionine was found. This number was used to quantitate all remaining spots in conjunction with the number of methionines present in the protein.

To ensure that proteins were labeled to equilibrium, parallel 2D gels were prepared and run on yeast metabolically labeled for 1, 2, 6, or 18 h. The corresponding 156 spots were excised from each gel, and radioactivity was measured by liquid scintillation counting for each spot. Calculated protein levels were highly reproducible for all time points measured after 1 h.

Calculation of codon bias and predicted half-life. Codon bias values were extracted from the YPD spreadsheet (17). Protein half-lives were calculated based on the N-end rule (33). When the N-terminal processing was not known experimentally, it was predicted based on the affinity of methionine aminopeptidase (31).

RESULTS

Characteristics of proteome approach. Nearly every facet of proteome analysis hinges on the unambiguous identification of large numbers of expressed proteins in cells. Several techniques have been described previously for the identification of proteins separated by 2DE, including N-terminal and internal sequencing (1, 2), amino acid analysis (38), and more recently mass spectrometry (25). We utilized techniques based on mass spectrometry because they afford the highest levels of sensitivity and provide unambiguous identification. The specific procedure used is schematically illustrated in Fig. 1 and is based on three principles. First, proteins are removed from the gel by proteolytic in-gel digestion, and the resulting peptides are separated by on-line capillary high-performance liquid chromatography. Second, the eluting peptides are ionized and detected, and the specific peptide ions are selected and fragmented by the mass spectrometer. To achieve this, the mass spectrometer switches between the MS mode (for peptide mass identification) and the MS/MS mode (for peptide characterization and sequencing). Selected peptides are fragmented by a process called collision-induced dissociation (CID) to generate a tandem mass spectrum (MS/MS spectrum) that contains the peptide sequence information. Third, individual CID mass spectra are then compared by computer algorithms to predicted spectra from a sequence database. This results in the identification of the peptide and, by association, the protein(s) in the spot. Unambiguous protein identification is attained in a single analysis by the detection of multiple peptides derived from the same protein.

Protein identification. Yeast total cell protein lysate (40 μg), metabolically labeled with [^{35}S]methionine, was electrophoretically separated by isoelectric focusing in the first dimension and by SDS–10% polyacrylamide gel electrophoresis in the second dimension. Proteins were visualized by silver staining and by autoradiography. Of the more than 1,000 proteins visible by silver staining, 156 spots were excised from the gel and subjected to in-gel tryptic digestion, and the resulting peptides were analyzed and identified by microspray LC-MS/MS techniques as described above. The proteins in this study were all identified automatically by computer software with no human interpretation of mass spectra. They are indicated in Fig. 2 and detailed in Table 1.

The CID spectra shown in Fig. 3 indicate that the quality of the identification data generated was suitable for unambiguous protein identification. The spectra represent the amino acid sequences of tryptic peptides NSGDIVNLGSIAGR (Fig. 3A) and FAVGAFTDSLR (Fig. 3B). Both peptides were derived from protein S57593 (hypothetical protein YMR226C), which migrated to spot 114 (molecular weight, 29,156; pI, 6.59) in the 2D gel in Fig. 2. Five other peptides from the same analysis were also computer matched to the same protein sequence.

Protein and mRNA quantitation. For the 156 genes investigated, the protein expression levels ranged from 2,200 (PGM2) to 863,000 (TDH2/TDH3) copies/cell. The levels of mRNA for each of the genes identified were calculated from SAGE frequency tables (35). These tables contain the mRNA levels for 4,665 genes in yeast strain YPH499 grown to mid-log phase in YPD medium on glucose as a carbon source. In some instances, the mRNA levels could not be calculated for reasons stated in Materials and Methods. For the proteins analyzed in this study, mean transcript levels varied from 0.7 to 473 copies/cell.

Selection of the sample population for mRNA-protein expression level correlation. The protein spots selected for identification were selected from spots visible by silver staining in the 2D gel. An attempt was made not to include spots where overlap with other spots was readily apparent. The number of proteins identified was 156 (Table 1). Some proteins migrated to more than one spot (presumably due to differential protein processing or modifications), and protein levels from these spots were calculated by integrating the intensities of the different spots. The 156 protein spots analyzed represented the products of 128 different genes. Genes were excluded from the correlation analysis only if part of the data set was missing; i.e., genes were excluded if (i) no mRNA expression data were available for the protein or putative SAGE tags were ambiguous, (ii) the amino acid sequence did not contain methionine, (iii) more than a single protein was conclusively identified as

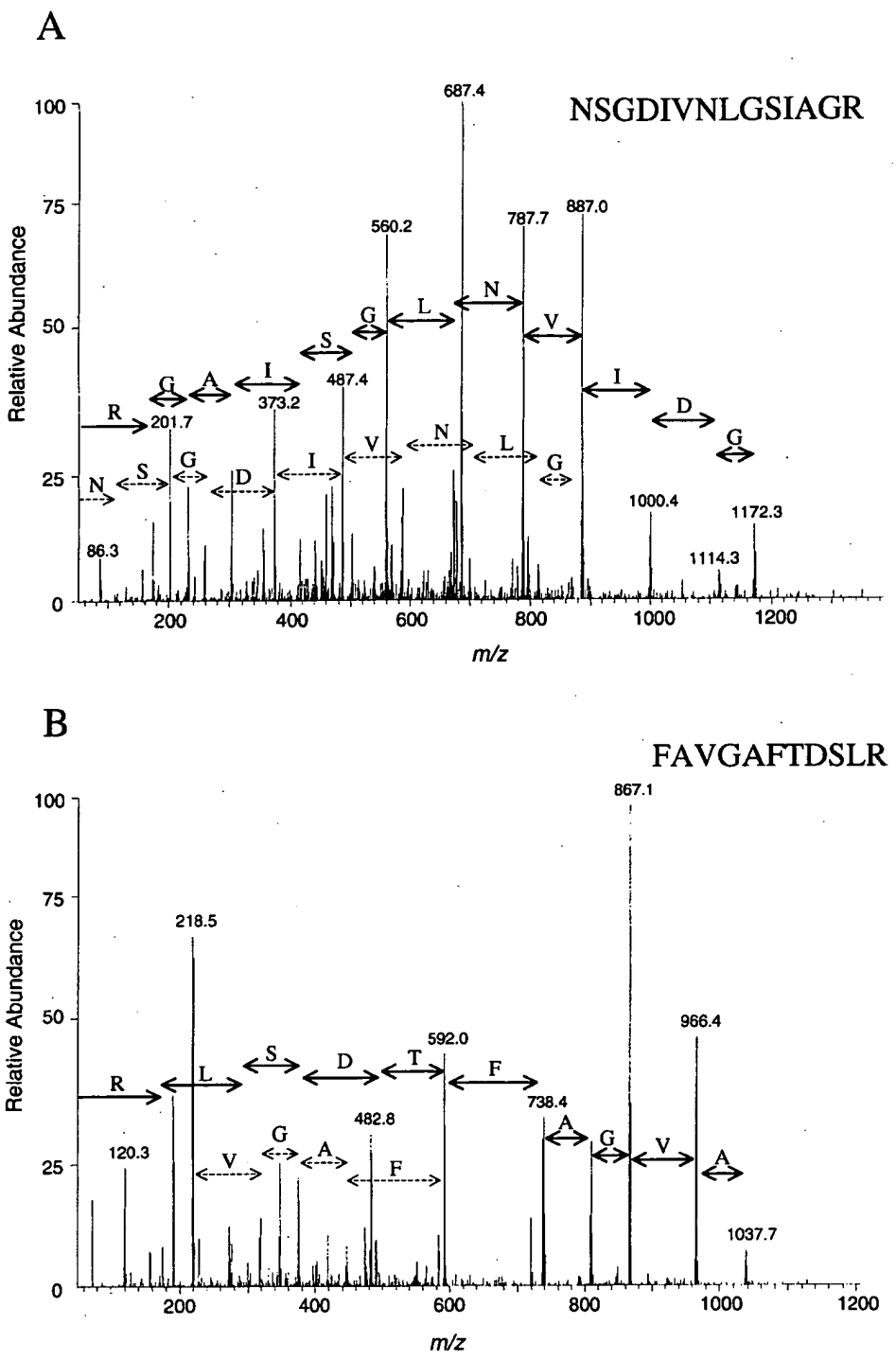


FIG. 3. Tandem mass (MS/MS) spectra resulting from analysis of a single spot on a 2D gel. The first quadrupole selected a single mass-to-charge ratio (*m/z*) of 687.2 (A) or 592.6 (B), while the collision cell was filled with argon gas, and a voltage which caused the peptide to undergo fragmentation by CID was applied. The third quadrupole scanned the mass range from 50 to 1,400 *m/z*. The computer program Sequest (8) was utilized to match MS/MS spectra to amino acid sequence by database searching. Both spectra matched peptides from the same protein, S57593 (yeast hypothetical protein YMR226C). Five other peptides from the same analysis were matched to the same protein.

migrating to the same gel spot, or (iv) the theoretical and observed pIs and molecular weights could not be reconciled. After these criteria were applied, the number of genes used in the correlation analysis was 106.

Codon bias and predicted half-lives. Codon bias is thought to be an indicator of protein expression, with highly expressed proteins having large codon bias values. The codon bias distribution for the entire set of more than 6,000 predicted yeast

gene ORFs is presented in Fig. 4A. The interval with the largest frequency of genes is between the codon bias values of 0.0 and 0.1. This segment contains more than 2,500 genes. The distribution of the codon bias values of the 128 different genes found in this study (all protein spots from Fig. 2) is shown in Fig. 4B, and protein half-lives (predicted from applying the N-end rule [33] to the experimentally determined or predicted protein N termini) are shown in Fig. 4C. No genes were identified with codon bias values less than 0.1 even though thousands of genes exist in this category. In addition, nearly all of the proteins identified had long predicted half-lives (greater than 30 h).

Correlation of mRNA and protein expression levels. The correlation between mRNA and protein levels of the genes selected as described above is shown in Fig. 5. For the entire group (106 genes) for which a complete data set was generated, there was a general trend of increased protein levels resulting from increased mRNA levels. The Pearson product moment correlation coefficient for the whole data set (106 genes) was 0.935. This number is highly biased by a small number of genes with very large protein and message levels. A more representative subset of the data is shown in the inset of Fig. 5. It shows genes for which the message level was below 10 copies/cell and includes 69% (73 of 106 genes) of the data used in the study. The Pearson product moment correlation coefficient for this data set was only 0.356. We also found that levels of protein expression coded for by mRNA with comparable abundance varied by as much as 30-fold and that the mRNA levels coding for proteins with comparable expression levels varied by as much as 20-fold.

The distortion of the correlation value induced by the uneven distribution of the data points along the *x* axis is further demonstrated by the analysis in Fig. 6. The 106 samples included in the study were ranked by protein abundance, and the Pearson product moment correlation coefficient was repeatedly calculated after including progressively more, and higher-abundance, proteins in each calculation. The correlation values remained relatively stable in the range of 0.1 to 0.4 if the lowest-expressed 40 to 95 proteins used in this study were included. However, the correlation value steadily climbed by the inclusion of each of the 11 very highly expressed proteins.

Correlation of protein and mRNA expression levels with codon bias. Codon bias is the propensity for a gene to utilize the same codon to encode an amino acid even though other codons would insert the identical amino acid in the growing polypeptide sequence. It is further thought that highly expressed proteins have large codon biases (3). To assess the value of codon bias for predicting mRNA and protein levels in exponentially growing yeast cells, we plotted the two experimental sets of data versus the codon bias (Fig. 7). The distribution patterns for both mRNA and protein levels with respect to codon bias were highly similar. There was high variability in the data within the codon bias range of 0.8 to 1.0. Although a large codon bias generally resulted in higher protein and message expression levels, codon bias did not appear to be predictive of either protein levels or mRNA levels in the cell.

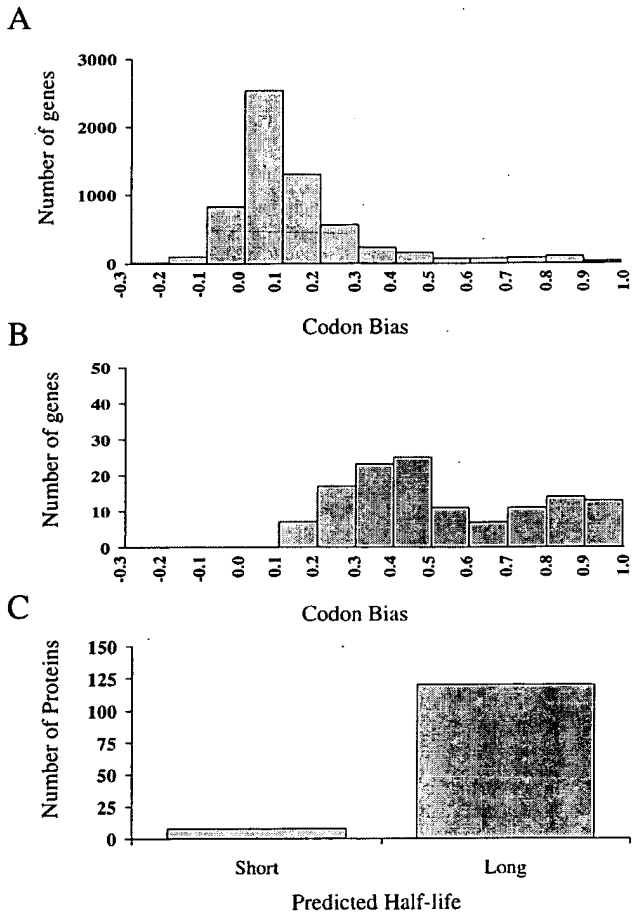


FIG. 4. Current proteome analysis technology utilizing 2DE without preenrichment samples mainly highly expressed and long-lived proteins. Genes encoding highly expressed proteins generally have large codon bias values. (A) Distribution of the yeast genome (more than 6,000 genes) based on codon bias. The interval with the largest frequency of genes is 0.0 to 0.1, with more than 2,500 genes. (B) Distribution of the genes from identified proteins in this study based on codon bias. No genes with codon bias values less than 0.1 were detected in this study. (C) Distribution of identified proteins in this study based on predicted half-life (estimated by N-end rule).

DISCUSSION

The desired end point for the description of a biological system is not the analysis of mRNA transcript levels alone but also the accurate measurement of protein expression levels and their respective activities. Quantitative analysis of global mRNA levels currently is a preferred method for the analysis of the state of cells and tissues (11). Several methods which either provide absolute mRNA abundance (34, 35) or relative mRNA levels in comparative analyses (20, 27) have been described elsewhere. The techniques are fast and exquisitely sensitive and can provide mRNA abundance for potentially any expressed gene. Measured mRNA levels are often implicitly or explicitly extrapolated to indicate the levels of activity of the corresponding protein in the cell. Quantitative analysis of protein expression levels (proteome analysis) is much more time-consuming because proteins are analyzed sequentially one by one and is not general because analyses are limited to the relatively highly expressed proteins. Proteome analysis does, however, provide types of data that are of critical importance for the description of the state of a biological system and that are not readily apparent from the sequence and the level of expression of the mRNA transcript. This study attempts to examine the relationship between mRNA and protein expression levels for a large number of expressed genes in cells representing the same state.

Limits in the sensitivity of current protein analysis technology precluded a completely random sampling of yeast proteins. We therefore based the study on those proteins visible by silver

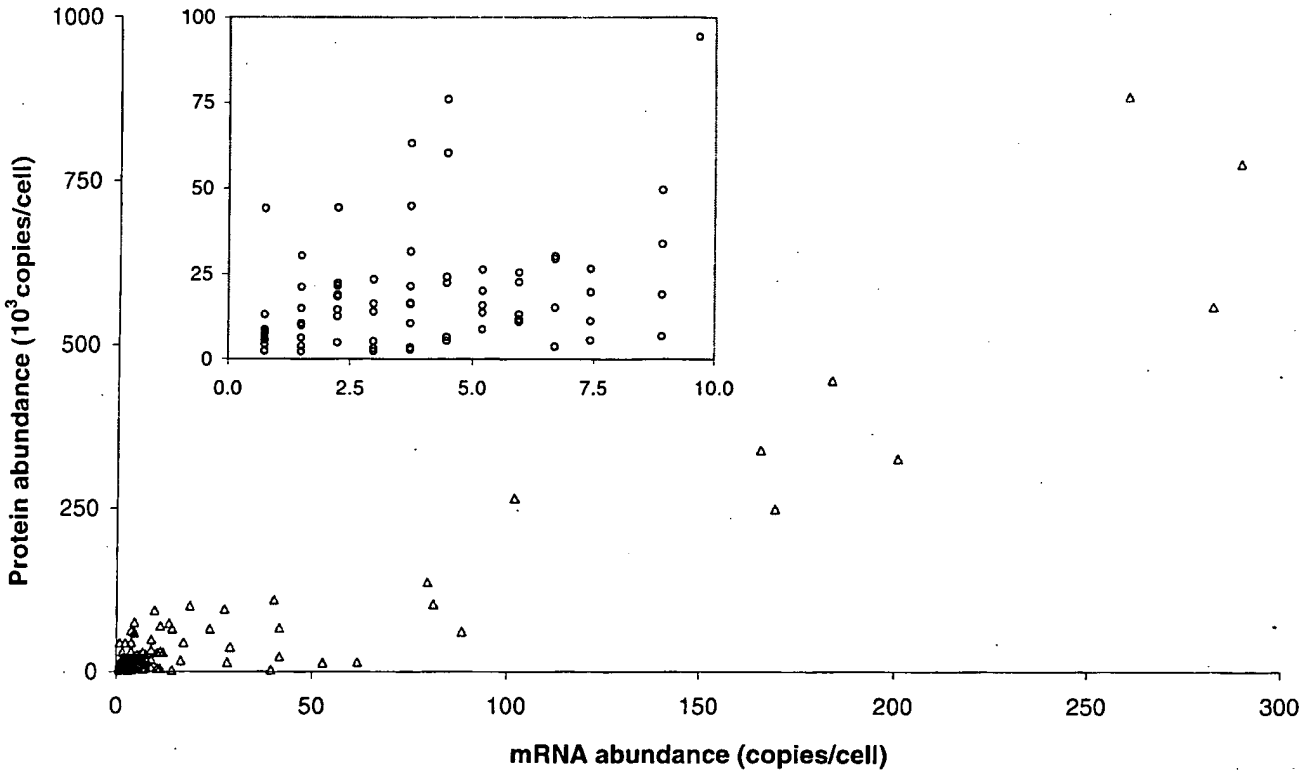


FIG. 5. Correlation between protein and mRNA levels for 106 genes in yeast growing at log phase with glucose as a carbon source. mRNA and protein levels were calculated as described in Materials and Methods. The data represent a population of genes with protein expression levels visible by silver staining on a 2D gel chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities. The inset shows the low-end portion of the main figure. It contains 69% of the original data set. The Pearson product moment correlation for the entire data set was 0.935. The correlation for the inset containing 73 proteins (69%) was only 0.356.

staining on a 2D gel. Of the more than 1,000 visible spots, 156 were chosen to include the entire range of molecular weights, isoelectric focusing points, and staining intensities displayed on the 2D protein pattern. The genes identified in this study shared a number of properties. First, all of the proteins in this study had a codon bias of greater than 0.1 and 93% were greater than 0.2 (Fig. 4B). Second, with few exceptions, the proteins in this study had long predicted half-lives according to the N-end rule (Fig. 4C). Third, low-abundance proteins with regulatory functions such as transcription factors or protein kinases were not identified.

Because the population of proteins used in this study appears to be fairly homogeneous with respect to predicted half-life and codon bias, it might be expected that the correlation of the mRNA and protein expression levels would be stronger for this population than for a random sample of yeast proteins. We tested this assumption by evaluating the correlation value if different subsets of the available data were included in the calculation. The 106 proteins were ranked from lowest to highest protein expression level, and the trend in the correlation value was evaluated by progressively including more of the higher-abundance proteins in the calculation (Fig. 6). The correlation value when only the lower-abundance 40 to 93 proteins were examined was consistently between 0.1 and 0.4. If the 11 most abundant proteins were included, the correlation steadily increased to 0.94. We therefore expect that the correlation for all yeast proteins or for a random selection would be less than 0.4. The observed level of correlation between mRNA and protein expression levels suggests the importance of posttranslational mechanisms controlling gene expression. Such mechanisms include translational control (15) and control of protein half-life (33). Since these mechanisms are also active in higher eukaryotic cells, we speculate that there is no predictive correlation between steady-state levels of mRNA and those of protein in mammalian cells.

Like other large-scale analyses, the present study has several potential sources of error related to the methods used to determine mRNA and protein expression levels. The mRNA levels were calculated from frequency tables of SAGE data. This method is highly quantitative because it is based on actual sequencing of unique tags from each gene, and the number of times that a tag is represented is proportional to the number of mRNA molecules for a specific gene. This method has some limitations including the following: (i) the magnitude of the error in the measurement of mRNA levels is inversely proportional to the mRNA levels, (ii) SAGE tags from highly similar genes may not be distinguished and therefore are summed, (iii) some SAGE tags are from sequences in the 3′ untranslated region of the transcript, (iv) incomplete cleavage at the SAGE tag site by the restriction enzyme can result in two tags representing one mRNA, and (v) some transcripts actually do not generate a SAGE tag (34, 35).

For the SAGE method, the error associated with a value increases with a decreasing number of transcripts per cell. The conclusions drawn from this study are dependent on the quality of the mRNA levels from previously published data (35). Since more than 65% of the mRNA levels included in this study were calculated to 10 copies/cell or less (40% were less

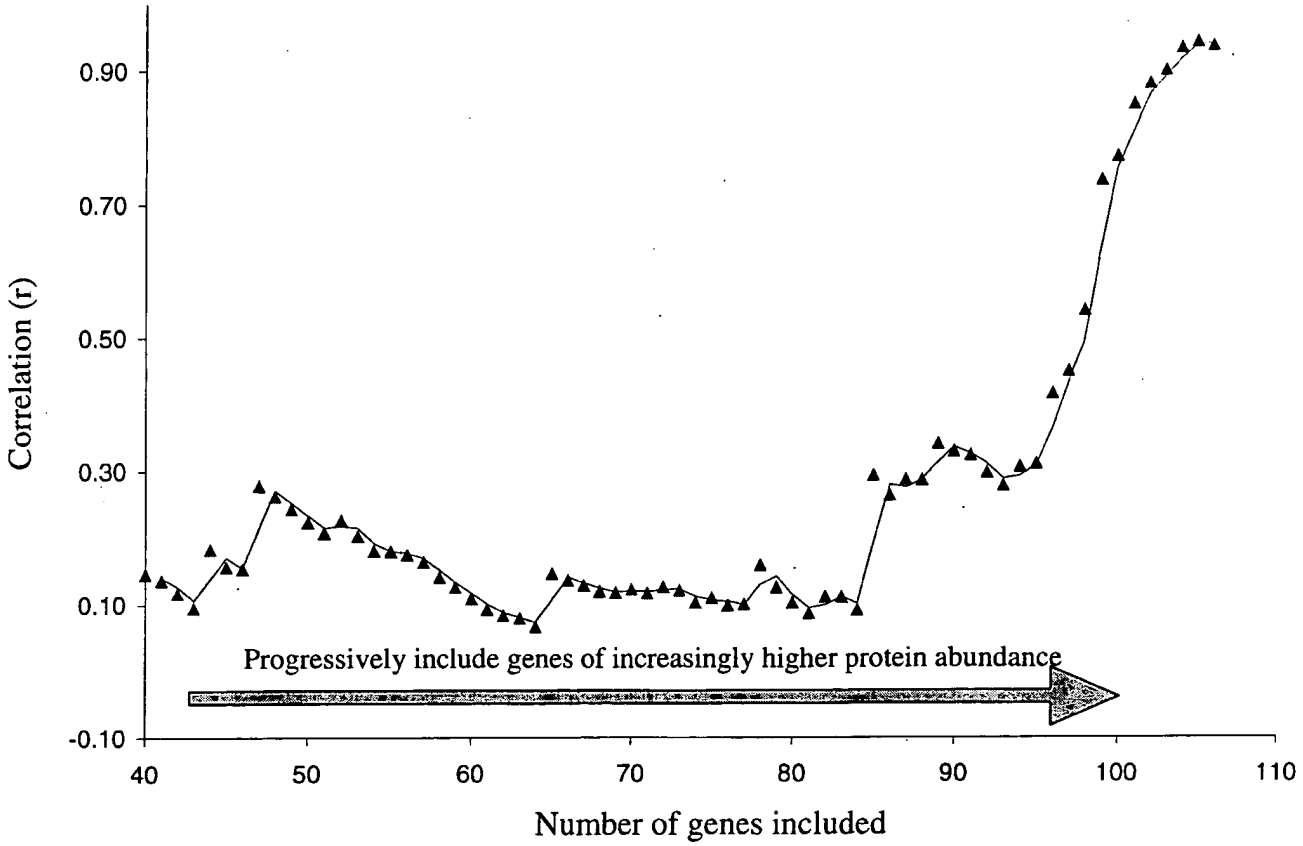


FIG. 6. Effect of highly abundant proteins on Pearson product moment correlation coefficient for mRNA and protein abundance in yeast. The set of 106 genes was ranked according to protein abundance, and the correlation value was calculated by including the 40 lowest-abundance genes and then progressively including the remaining 66 genes in order of abundance. The correlation value climbs as the final 11 highly abundant proteins are included.

than 4 copies/cell), the error associated with these values may be quite large. The mRNA levels were calculated from more than 20,000 transcripts. Assuming that the estimate of 15,000 mRNA molecules per cell is correct (16), this would mean that mRNA transcripts present at only a single copy per cell would be detected 72% of the time (35). The mRNA levels for each gene were carefully scrutinized, and only mRNA levels for which a high degree of confidence existed were included in the correlation value.

Protein abundance was determined by metabolic radiolabeling with [^{35}S]methionine. The calculation required knowledge of three variables: the number of methionines in the mature protein, the radioactivity contained in the protein, and the specific activity of the radiolabel normalized per methionine. The number of methionines per protein was determined from the amino acid sequence of the proteins identified by tandem mass spectrometry. For some proteins, it was not known whether the methionine of the nascent polypeptide was processed away. The N termini of those proteins were predicted based on the specificity of methionine aminopeptidase (31). If the N-terminal processing did not conform to the predicted specificity of processing enzymes, the calculation of the number of methionines would be affected. This discrepancy would affect most the quantitation of a protein with a very low number of methionines. The average number of calculated methionines per protein in this study was 7.2. We therefore expect the potential for erroneous protein quantitation due to unusual N-terminal processing to be small.

The amount of radioactivity contained in a single spot might be the sum of the radioactivity of comigrating proteins. Because protein identification was based on tandem mass spectrometric techniques, comigrating proteins could be identified. However, comigrating proteins were rarely detected in this study, most likely because relatively small amounts of total protein (40 μg) were initially loaded onto the gels, which resulted in highly focused spots containing generally 1 to 25 ng of protein. Because of the relatively small amount loaded, the concentrations of any potentially comigrating protein would likely be below the limit of detection of the mass spectrometry technique used in this study (1 to 5 ng) and below the limit of visualization by silver staining (1 to 5 ng). In the overwhelming majority of the samples analyzed, numerous peptides from a single protein were detected. It is assumed that any comigrating proteins were at levels too low to be detected and that their influence in the calculation would be small.

The specific activity of the radiolabel was determined by relating the precise amount of protein present in selected spots of a parallel gel, as determined by quantitative amino acid composition analysis, to the number of methionines present in the sequence of those proteins and the radioactivity determined by liquid scintillation counting. It is possible that the resulting number might be influenced by unavoidable losses inherent in the amino acid analysis procedure applied. Because four different proteins were utilized in the calculation and the experiment was done in duplicate, the specific activity calculated is thought to be highly accurate. Indeed, the specific

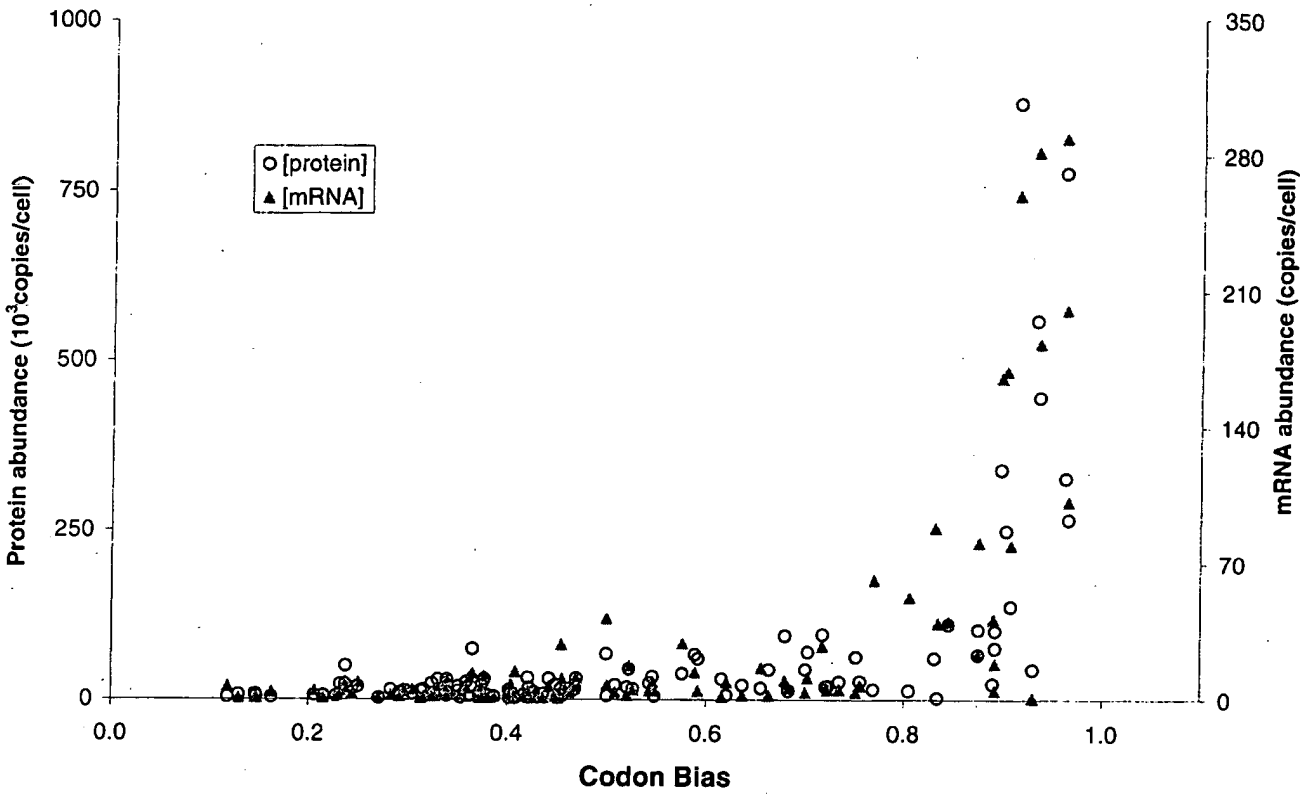


FIG. 7. Relationship between codon bias and protein and mRNA levels in this study. Yeast mRNA and protein expression levels were calculated as described in Materials and Methods. The data represent the same 106 genes as in Fig. 5.

activities calculated for each of the four proteins varied by less than 10%. Any inconsistencies in the calculation of the specific activity would result in differences in the absolute levels calculated but not in the relative numbers and would therefore not influence the correlation value determined.

The protein quantitative method used eliminates a number of potential errors inherent in previous methods for the quantitation of proteins separated by 2DE, such as preferential protein staining and bias caused by inequalities in the number of radiolabeled residues per protein. Any 2D gel-based method of quantitation is complicated by the fact that in some cases the translation products of the same mRNA migrated to different spots. One major reason is posttranslational modification or processing of the protein. Also, artifactual proteolysis during cell lysis and sample preparation can lead to multiple resolved forms of the protein. In such cases, the protein levels of spots coded for by the same mRNA were pooled. In addition, the existence of other spots coded for by the same mRNA that were not analyzed by mass spectrometry or that were below the limit of detection for silver staining cannot be ruled out. However, since this study is based on a class of highly expressed proteins, the presence of undetected minor spots below silver staining sensitivity corresponding to a protein analyzed in the study would generally cause a relatively small error in protein quantitation.

Codon bias is a measure of the propensity of an organism to selectively utilize certain codons which result in the incorporation of the same amino acid residue in a growing polypeptide chain. There are 61 possible codons that code for 20 amino acids. The larger the codon bias value, the smaller the number of codons that are used to encode the protein (19). It is thought that codon bias is a measure of protein abundance because highly expressed proteins generally have large codon bias values (3, 13).

Nearly all of the most highly expressed proteins had codon bias values of greater than 0.8. However, we detected a number of genes with high codon bias and relative low protein abundance (Fig. 7). For example, the expressed gene with both the second largest protein and mRNA levels in the study was ENO2_YEAST (775,000 and 289.1 copies/cell, respectively). ENO1_YEAST was also present in the gel at much lower protein and mRNA levels (44,200 and 0.7 copies/cell, respectively). The codon bias values for ENO2 and ENO1 are similar (0.96 and 0.93, respectively), but the expression of the two genes is differentially regulated. Specifically, ENO1_YEAST is glucose repressed (6) and was therefore present in low abundance under the conditions used. Other genes with large codon bias values that were not of high protein abundance in the gel include EFT1, TIF1, HXK2, GSP1, EGD2, SHM2, and TAL1. We conclude that merely determining the codon bias of a gene is not sufficient to predict its protein expression level.

Interestingly, codon bias appears to be an excellent indicator of the boundaries of current 2D gel proteome analysis technology. There are thousands of genes with expressed mRNA and likely expressed protein with codon bias values less than 0.1 (Fig. 4A). In this study, we detected none of them, and only a very small percentage of the genes detected in this study had codon bias values between 0.1 and 0.2 (Fig. 4B). Indeed, in every examined yeast proteome study (5, 7, 13, 28) where the combined total number of identified proteins is 300 to 400, this same observation is true. It is expected that for the more complex cells of higher eukaryotic organisms the detection of

low-abundance proteins would be even more challenging than for yeast. This indicates that highly abundant, long-lived proteins are overwhelmingly detected in proteome studies. If proteome analysis is to provide truly meaningful information about cellular processes, it must be able to penetrate to the level of regulatory proteins, including transcription factors and protein kinases. A promising approach is the use of narrow-range focusing gels with immobilized pH gradients (IPG) (23). This would allow for the loading of significantly more protein per pH unit covered and also provide increased resolution of proteins with similar electrophoretic mobilities. A standard pH gradient in an isoelectric focusing gel covers a 7-pH-unit range (pH 3 to 10) over 18 cm. A narrow-range focusing gel might expand the range to 0.5 pH units over 18 cm or more. This could potentially increase by more than 10-fold the number of proteins that can be detected. Clearly, current proteome technology is incapable of analyzing low-abundance regulatory proteins without employing an enrichment method for relatively low-abundance proteins. In conclusion, this study examined the relationship between yeast protein and message levels and revealed that transcript levels provide little predictive value with respect to the extent of protein expression.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Science and Technology Center for Molecular Biotechnology, NIH grant T32HG00035-3, and a grant from Oxford Glycosciences.

We thank Jimmy Eng for expert computer programming, Garry Corthals and John R. Yates III for critical discussion, and Siavash Mohandesi for expert technical help.

REFERENCES

1. **Aebersold, R. H., J. Leavitt, R. A. Saavedra, L. E. Hood, and S. B. Kent.** 1987. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. Proc. Natl. Acad. Sci. USA **84**:6970–6974.
2. **Aebersold, R. H., D. B. Teplow, L. E. Hood, and S. B. Kent.** 1986. Electroblotting onto activated glass. High efficiency preparation of proteins from analytical sodium dodecyl sulfate-polyacrylamide gels for direct sequence analysis. Eur. J. Biochem. **261**:4229–4238.
3. **Bennetzen, J. L., and B. D. Hall.** 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.
4. **Boucherie, H., G. Dujardin, M. Kermorgant, C. Monribot, P. Slonimski, and M. Perrot.** 1995. Two-dimensional protein map of Saccharomyces cerevisiae: construction of a gene-protein index. Yeast **11**:601–613.
5. **Boucherie, H., F. Sagliocco, R. Joubert, I. Maillet, J. Labarre, and M. Perrot.** 1996. Two-dimensional gel protein database of Saccharomyces cerevisiae. Electrophoresis **17**:1683–1699.
6. **Carmen, A. A., P. K. Brindle, C. S. Park, and M. J. Holland.** 1995. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene ENO1. Yeast **11**:1031–1043.
7. **Ducret, A., I. VanOostveen, J. K. Eng, J. R. Yates, and R. Aebersold.** 1998. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. Protein Sci. **7**:706–719.
8. **Eng, J., A. McCormack, and J. R. Yates.** 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5**:976–989.
9. **Figeys, D., A. Ducret, J. R. Yates, and R. Aebersold.** 1996. Protein identification by solid phase microextraction-capillary zone electrophoresis-microelectrospray-tandem mass spectrometry. Nat. Biotechnol. **14**:1579–1583.
10. **Figeys, D., I. VanOostveen, A. Ducret, and R. Aebersold.** 1996. Protein identification by capillary zone electrophoresis/microelectrospray ionization-tandem mass spectrometry at the subfemtomole level. Anal. Chem. **68**:1822–1828.
11. **Fraser, C. M., and R. D. Fleischmann.** 1997. Strategies for whole microbial genome sequencing and analysis. Electrophoresis **18**:1207–1216.
12. **Garrels, J. I., B. Futcher, R. Kobayashi, G. I. Latter, B. Schwender, T. Volpe, J. R. Warner, and C. S. McLaughlin.** 1994. Protein identifications for a Saccharomyces cerevisiae protein database. Electrophoresis **15**:1466–1486.
13. **Garrels, J. I., C. S. McLaughlin, J. R. Warner, B. Futcher, G. I. Latter, R. Kobayashi, B. Schwender, T. Volpe, D. S. Anderson, F. Mesquita-Fuentes, and W. E. Payne.** 1997. Proteome studies of Saccharomyces cerevisiae: identification and characterization of abundant proteins. Electrophoresis **18**: 1347–1360.
14. **Gygi, S. P., and R. Aebersold.** 1998. Absolute quantitation of 2-DE protein spots, p. 417–421. *In* A. J. Link (ed.), 2-D protocols for proteome analysis. Humana Press, Totowa, N.J.
15. **Harford, J. B., and D. R. Morris.** 1997. Post-transcriptional gene regulation. Wiley-Liss, Inc., New York, N.Y.
16. **Hereford, L. M., and M. Rosbash.** 1977. Number and distribution of polyadenylated RNA sequences in yeast. Cell **10**:453–462.
17. **Hodges, P. E., W. E. Payne, and J. I. Garrels.** 1998. The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae. Nucleic Acids Res. **26**:68–72.
18. **Klose, J., and U. Kobalz.** 1995. Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. Electrophoresis **16**:1034–1059.
19. **Kurland, C. G.** 1991. Codon bias and gene expression. FEBS Lett. **285**:165–169.
20. **Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis.** 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA **94**:13057–13062.
21. **Liang, P., and A. B. Pardee.** 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science **257**:967–971.
22. **Link, A. J., L. G. Hays, E. B. Carmack, and J. R. Yates III.** 1997. Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143. Electrophoresis **18**:1314–1334.
23. **Nawrocki, A., M. R. Larsen, A. V. Podtelejnikov, O. N. Jensen, M. Mann, P. Roepstorff, A. Gorg, S. J. Fey, and P. M. Larsen.** 1998. Correlation of acidic and basic carrier ampholyte and immobilized pH gradient two-dimensional gel electrophoresis patterns based on mass spectrometric protein identification. Electrophoresis **19**:1024–1035.
24. **O'Farrell, P. H.** 1975. High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. **250**:4007–4021.

24a. **OWL Protein Sequence Database.** 2 August 1998, posting date. [Online.] http://bmbsgi11.leeds.ac.uk/bmb5dp/owl.html. [8 January 1999, last date accessed.]

25. **Patterson, S. D., and R. Aebersold.** 1995. Mass spectrometric approaches for the identification of gel-separated proteins. Electrophoresis **16**:1791–1814.
26. **Pennington, S. R., M. R. Wilkins, D. F. Hochstrasser, and M. J. Dunn.** 1997. Proteome analysis: from protein characterization to biological function. Trends Cell Biol. **7**:168–173.
27. **Shalon, D., S. J. Smith, and P. O. Brown.** 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. **6**:639–645.
28. **Shevchenko, A., O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, and M. Mann.** 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc. Natl. Acad. Sci. USA **93**:14440–14445.
29. **Shevchenko, A., M. Wilm, O. Vorm, and M. Mann.** 1996. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. Anal. Chem. **68**:850–858.
30. **Sikorski, R. S., and P. Hieter.** 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics **122**:19–27.
31. **Tsunasawa, S., J. W. Stewart, and F. Sherman.** 1985. Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. J. Biol. Chem. **260**:5382–5391.
32. **Urlinger, S., K. Kuchler, T. H. Meyer, S. Uebel, and R. Tamp'e.** 1997. Intracellular location, complex formation, and function of the transporter associated with antigen processing in yeast. Eur. J. Biochem. **245**:266–272.
33. **Varshavsky, A.** 1996. The N-end rule: functions, mysteries, uses. Proc. Natl. Acad. Sci. USA **93**:12142–12149.
34. **Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler.** 1995. Serial analysis of gene expression. Science **270**:484–487.
35. **Velculescu, V. E., L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, Jr., P. Hieter, B. Vogelstein, and K. W. Kinzler.** 1997. Characterization of the yeast transcriptome. Cell **88**:243–251.
36. **Wilkins, M. R., K. L. Williams, R. D. Appel, and D. F. Hochstrasser.** 1997. Proteome research: new frontiers in functional genomics. Springer-Verlag, Berlin, Germany.
37. **Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, and M. Mann.** 1996. Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. Nature **379**:466–469.
38. **Yan, J. X., M. R. Wilkins, K. Ou, A. A. Gooley, K. L. Williams, J. C. Sanchez, O. Golaz, C. Pasquali, and D. F. Hochstrasser.** 1996. Large-scale amino-acid analysis for proteome studies. J. Chromatogr. A **736**:291–302.
39. **YPD Website.** 6 March 1998, revision date. [Online.] Proteome, Inc. http://www.proteome.com/YPDhome.html. [8 January 1999, last date accessed.]